CHAPTER 27

A Digital Humanities Approach to the History of Culture and Science
Drugs and Eugenics Revisited in Early 20th-Century Dutch Newspapers, Using Semantic Text Mining

Stephen Snelders[a], Pim Huijnen[a,b], Jaap Verheul[b], Maarten de Rijke[c] and Toine Pieters[a]

[a]Descartes Centre for the History and Philosophy of the Sciences and the Arts, Freudenthal Institute, Utrecht University(UU), The Netherlands {s.snelders, t.pieters@uu.nl},
[b]Research Institute for History and Art History, UU {p.huijnen, j.verheul@uu.nl}
[c]ISLA, University of Amsterdam, The Netherlands {m.derijke@uva.nl}

ABSTRACT

Human language technology developed and used in CLARIN demonstrator projects WAHSP and BILAND supports advanced forms of (multi-lingual) text mining of large datasets of newspapers. We argue that the combination of exploratory search and text mining offers an innovative research approach to systematically set up search trails in the historical sciences. We describe the development, use, and methodological challenges of the WAHSP and BILAND text-mining tools and the successor tool, Texcavator, to support alternating forms of distant reading and close reading in newspaper collections. We will show how semantic text mining speeds up the heuristic process and thus helped to provide new and challenging perspectives on the circulation of ideas and notions regarding drugs and eugenics in Dutch newspapers in the first four decades of the 20th century.

27.1 Introduction

Historical scholars are increasingly applying computational tools and methods to all phases of their research. Digital tools are used to open, present, and curate textual and multi-media sources in semantic text mining, for integration of geospatial information data, for various forms of visualisation, and for enhanced and multi-media publication of research results, blogs, and wikis. Digital history is a methodological approach that is framed by these digital tools' ability to make, define,

How to cite this book chapter:
Snelders, S, Huijnen, P, Verheul, J, de Rijke, M and Pieters. T. 2017. A Digital Humanities Approach to the History of Culture and Science: Drugs and Eugenics Revisited in Early 20th-Century Dutch Newspapers, Using Semantic Text Mining. In: Odijk, J and van Hessen, A. (eds.) *CLARIN in the Low Countries*, Pp. 325–336. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbi.27.

query, and annotate associations and explore long-term patterns of economic, technological, and cultural change in past human records. Digital history touches on all aspects and forms of historical scholarship that come together around digitised data and digital tools (Graham et al., 2013; Van Eijnatten et al., 2013). There have been inspiring examples of good digital historical scholarship, like the use of N-grams to mine the Google Books archive (Michel, 2010), the engagement in creative visual analysis of historical geography,[1] or the study of the 'circulation of knowledge and learned practices' by means of a virtual research environment (VRE).[2] However, historians are still in the process of learning how to incorporate and implement data-mining technology in methodologically sound and reproducible ways (Seefeldt and Thomas, 2009; Bingham, 2010; Earheart and Jewell, 2011; Berry, 2012; Burdick et al., 2012; Van Eijnatten et al., 2013; Van Eijnatten et al., 2014).[3]

Semantic text analytics is a particularly promising form of text mining that can be applied to 'big data' sets. Text analytics, or text mining, is an umbrella term for the incorporation and implemention of a wide range of tools or techniques (algorithms, statistics), including data mining, machine learning, natural language processing, and artificial intelligence (Jackson and Moulinier, 2007). The goal of text mining is to reduce the effort required to obtain meaningful information from large digitised text data sources. In principle, text-mining tools can process large numbers of texts reasonably quickly and support researchers in tracing sentiments, potentially meaningful events, and context-related concepts. However, being able to retrieve historically meaningful information requires that historians as domain users have a prominent role in the development of text- and data-mining technology.

Research programmes such as Digging into Data,[4] CLARIN-NL[5] and CLARIAH[6] demonstrate the feasibility of performing interdisciplinary humanities research facilitated by digital research tools. These programmes also show that collaborative, interdisciplinary and integrative strategies such as common group learning (where all knowledge is necessarily pooled, and learning is both shared and cumulative), modelling, and iterative and incremental approaches are central to the function, and, therefore, success, of digital humanities. It is therefore important to include articulating and aligning user needs, in our case historians, with technological options. For instance, incorporation of regular feedback loops allows for an iterative refinement of text-mining algorithms (e.g. identifying polarities and named entity recognition, etc.) and the development of a user-friendly interface (Warwick et al., 2012; Huijnen et al., 2014).

The combination of exploratory search and text mining has supported our research team to set up systematic search trails. Our thesis is that this approach enables fruitful alternating modes of distant reading and close reading. We will demonstrate our thesis by presenting case studies of debates about drugs and eugenics. Both topics represent a meeting ground between science and society with shifting cultural and political connotations. Drug use and eugenics as controversial social practices represent not only an important component of our cultural heritage, but also key elements of European modernisation (Hahn, 2000; Lombardo, 2001; Snelders et al., 2006; Reulecke, 2007; Levine and Bashford, 2010; Turda, 2010).

How can we use digitised newspaper collections to analyse the public's perceptions, opinions, and sentiments about eugenics and drugs? One approach suggests perusing lead and opinion articles about drugs or specific eugenic policy measures. The digital history projects discussed here have not limited their searches to this obvious choice. Rather, public debates are perceived as part

1 http://web.stanford.edu/group/spatialhistory/cgi-bin/site/index.php (accessed 03-02-2017)
2 http://ckcc.huygens.knaw.nl/ (accessed 25-01-2016)
3 For an overview of more recent trends and discussions in digital humanities see http://dh2016.adho.org/ (accessed 03-02-2017).
4 http://diggingintodata.org/ (accessed 03-02-2017)
5 http://clarin.nl/ (accessed 05-02-2016)
6 http://clariah.nl/ (accessed 05-02-2016)

of a public sphere in which perceptions, opinions, and sentiments are constructed and structured. The conceptual developments in the history of culture and science and their related disciplines after Foucault's discourse analysis suggest that implicit assumptions and perceptions of drugs and eugenics are pervasive throughout Western culture, and can be found in detective stories, advertisements, visual representations, journalistic reports and a wide range of other cultural texts. Implicit assumptions are a powerful expression of public perceptions and notions. For uninformed or biased readers, these 'hidden' discourses can construct and reinforce a range of associations of drug use as dangerous, criminal, anti-social and exotic, and the same is true for implicit associations with eugenics, in terms of deviant, weak and depraved social groups. We will show that the identification and analysis of these 'hidden' discourses and the possible overlap between discourses on drugs and eugenics generate potentially transformative insights into the construction and shift of meaning around science-related social practices such as drugs and eugenics.

27.2 Towards Historical Text Mining of Public Media

27.2.1 WAHSP Tool Features

The development of an open-source mining technology that historians without specific computer skills can and will use requires a user-friendly and user-informed interface. This was the basis requirement for developing the CLARIN-supported web application for historical-sentiment mining (a form of semantic text analytics that focuses on historical opinions, attitudes, and value judgements) of public media known as WAHSP. WAHSP was specifically designed for text mining the digital newspaper archive of the National Library of the Netherlands (Delpher collection). At present, this repository includes over 11 million pages from more than 200 newspapers and periodicals published between 1618 and 1995, which adds up to over 100 million articles.[7] WAHSP's technical basis is an ElasticSearch instance combined with the xTAS text analytics platform developed by the ISLA Informatics Institute of the University of Amsterdam.[8] xTAS includes modules for online and offline processing, and provides essential text pre-processing modules (morphological normalisation, format and encoding reconciliation, named-entity recognition and normalisation; Meij et al., 2009). It also incorporates algorithms and tools for the identification of polarity (positive/support or negative/criticism), sources (opinion-holders), frequency of items, and specific targets of discourses (Jijkoun et al., 2010). WAHSP comes with visualisation modules built in D3.js (interactive word clouds and timelines). WAHSP has been developed in a specific research context, but is generic and usable in other domains for which topic, context and attitude analysis is needed for large volumes of text.

The main added value of the WAHSP tool lies in exploratory readings of historical patterns in public debates. The WAHSP research team has found that in terms of methodology, semi-automatic document selection fits rather well with historical research as an alternative to manual or random sampling. This approach speeds up the heuristic process considerably. Word clouds that depict a linguistic context within which keywords occur are instrumental for helping an historian (with expert knowledge of the domain) to combine and compare various historical periods in a free associative manner on the basis of a large number of historical documents. Each query immediately yields a document selection without laborious sampling. Exploring word associations and metadata, as well as visualisations of the original newspaper articles over time, can lead to improved queries and allow the historian to alternate between distant and close reading and to systematically explore search trails.

---

[7] http://www.delpher.nl/nl/platform/pages/helpitems?nid=385 (accessed 03-02-2017)

[8] http://xtas.net/ (accessed 05-02-2016)

27.2.2 Exploring the Construction of Public Images of Drugs and Drug Users

Our aim was to analyse how Dutch newspapers represented debates on drugs, drug trafficking, and drug users in the early 20th century (1900–1940). We wanted to determine whether these early media debates on drugs were predominantly based on medical aspects (addiction, therapeutic benefit) or social aspects (crime, stigmatised groups of drug users).

How did we use WAHSP? First, we created a lexicon of terms related to drugs to capture all possible relevant terms for high document recall. We used our drug history domain knowledge to create a list of words, and WAHSP provided query-guided word (frequency) clouds based on all retrieved documents from the Delpher collection (see Figure 27.1). The word clouds enabled us to gradually expand the original query with terms we recognised as drug-related terms. We kept lab logs of all our queries and word cloud visualisation results.

Our point of departure was marking key events, such as the Shanghai Opium Conference (1909) and additional treaties, and the introduction and subsequent tightening of the Dutch Opium laws (1920 and 1928). We looked for word associations with drugs targeted by the Opium Law of 1919: opium, morphine, heroin and cocaine. We split these associations in four time periods: (I) before The Hague conference of 1912; (II) from 1912 to the enforcement of the Opium Law in 1920; (III) between 1920 and 1928, at which latter date the law was changed to make opium possession an offence; and (IV) from 1928 to the Second World War, when war-related issues took precedence in the mind of the public. It is significant to note that *after* 1920, the public's ideas about drugs seemed to change. The generic term *drugs* did not exist in the language of Dutch newspapers during the interwar period, but rather the terms were *narcotica* or *verdovende* or *verdoovende middelen* [all translated as 'narcotics']. Before 1920, newspapers used these Dutch words to refer to opium or cocaine (or chloroform, which had an important role as a narcotic in medicine). The number of articles addressing narcotics in this early period was quite limited compared to 1920–1928 and 1928–1940. After 1920, the number of articles increased by a factor of 16. The associations with these three words changed as well, from 'medicines', 'poisons', 'science', 'pharmacies', 'sleep' or 'narcosis' to 'police', 'contraband trade', 'arrested', and 'confiscated'. Not only are there increasing associations between the generic terms for narcotics and crime, but also for individual drugs and crime (Snelders and Pieters, 2012).

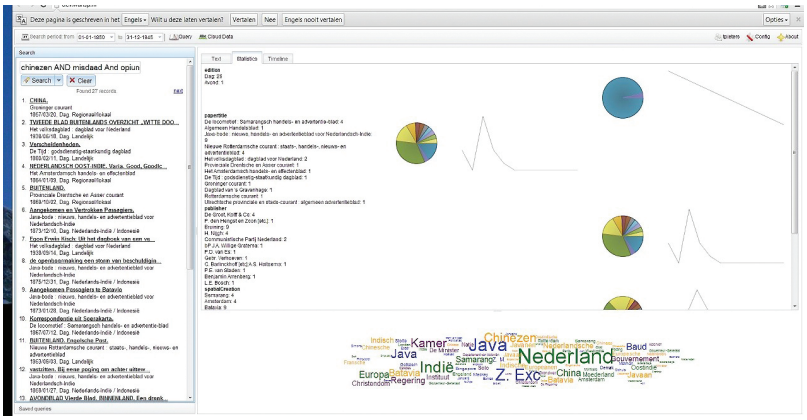

Figure 27.1: WAHSP search result plus word cloud based on the National Library of the Netherlands newspaper repository using the query 'chinezen' AND 'misdaad' AND 'opium' for 1900–1945. (6 September 2013).

By carefully inspecting word counts, we found exciting quantitative evidence for a historical caesura indicating that there was a criminalisation of the drug debate in approximately 1924 (Odijk et al., 2012). At that time, opium became associated with Chinese, China, or with the Dutch East Indies (see Figure 27.1). Even *Opiumregie*, the opium distribution and control regime of the colonial state in the Dutch East Indies, was initially associated with Chinese (and 'pathetic') users, and with Chinese crime syndicates that tried to evade the *Opiumregie* by smuggling illicit opiates into the East Indies. These opiates were smoked opium with a different quality than standard government-issued opium. In the 1930s, smugglers also attempted to bring morphine and heroin into the East Indies.

The negative and alienating ('othering') images of Chinese drug users and traders were found in the so-called 'hidden discourses' we mentioned above, which can also be associated with deviant groups and eugenics. We subsequently alternated from distant reading by means of exploratory search and text mining to conventional close reading of *Dick Bos*, a popular Dutch pulp comic series in the early 1940s, the first five parts of which were published in instalments in weekly papers from 1940–1942. The character Dick Bos was a detective who was an expert in an Asian fighting art (jiu jitsu). In his first appearance, he was immediately caught up in a drug case. Bos went on the trail of a gang smuggling cocaine on a Chinese ship. When the gang captured him, he was sent to China and put to work as a slave on a plantation. Of course, he escaped and ultimately helped to capture the gang.[9] That the actual relationship between cocaine and China was more complicated did not matter in this story; many (young) readers would have become firmly convinced of a connection between China, cocaine and deviancy. In *Dick Bos*, opium and cocaine were clearly and visually associated with a dark underworld, low-life taverns, and exotic and criminal Chinese groups. As such the alternation between distant and close reading provides evidence of a meaningful overlap and interference between drugs and eugenics discourses that deserves further research.

27.2.3 Mining Ambiguities in the Meaning of Eugenics

The WAHSP tool enables, as we have seen in the previous drugs case history, searches in the Delpher collection with combinations of keywords that do not necessarily refer specifically to eugenics, but rather imply eugenic thinking, such as in the case of the combination of 'Chinese' and crime. 'Eugenics' is a term loaded with historical meanings and polarities. Its literal meaning – 'good birth' – suggests a suitable goal for all prospective parents, yet its historical connotations tie it to a rather wide range of beliefs and practices, from good nutrition and education, pre-natal care for mothers, and birth-control to the extremes of selective breeding programmes, forced sterilisation, and euthanasia (Klausen and Bashford, 2010). Possible linguistic associations with eugenics include: 'ancestry', 'lineage', 'descent', 'reproduction', 'selection', 'unhealthy', 'pure'/'purity', 'weak', and 'deviant'.

By combining these words with keywords from social or cultural domains like sports, entertainment, economy and religion, one can obtain explicit discussions not only about eugenics, but also about implicit notions influenced by hereditary and eugenic thinking within certain debates. At the same time, one has to be aware that keyword searching is in itself not without problems – it is a rather 'blunt' instrument in the words of Adrian Bingham (Bingham, 2010: 229). Finding the right keywords demands expert knowledge of the field of study and significant perseverance and creativity (Nicholson, 2013: 67).

The hints of eugenic notions in pre-war Dutch economic debates can serve as yet another example that is highly suitable to illustrate what we mean by exploratory search methods. The economic

[9] Alfred Mazure, 'Het geval "Kleyn" in: *Dick Bos. Alle avonturen*, I. ('s-Gravenhage: Panda, 2005).

historian Thomas C. Leonard argues that economists in the Progressive Era (ca 1890–1920s) of the United States advocated for a minimum wage as a eugenic tool: a minimum wage would cause job losses and thus discourage prospective immigrants from coming to the US, as well as remove the more unfit (the so-called 'low wage races') from employment (Leonard, 2005: 213). It is an interesting question to consider whether similar arguments were used in Dutch debates on minimum wage, since although the Netherlands did not adopt a general minimum wage before 1968 the introduction of a minimum wage was debated from as early as the turn of the 20th century. The WAHSP tool generated almost 10,000 hits on 'minimum wage' before 1945.

The words *Amerika* and *Amerikaansche* [America and American] both form part of the resulting word cloud, indicating that the Dutch debate on the minimum wage might be informed by eugenic arguments from abroad – notably from the US. There are several exploratory angles to follow this trail. For instance, one can query (combinations of) relevant keywords that characterise this particular debate ('race', '(minimum) wage', 'immigration' and the like) or look for a possible link with the US. The combination 'race AND immigration'(see Figure 27.2), for example, hints at a connection with the US. Moreover, both the combinations 'wage AND race' and 'wage AND immigration' yielded a relatively high number of hits (more than 9,000 and almost 2,500, respectively; Huijnen et al., 2014: 78)

We operationalised the combination of exploratory search and text-mining as the ongoing process of refining and expanding queries with the help of text-mining techniques. Word clouds are an important part of this digital search method, because striking words can trigger the historian who has a profound knowledge of the subject matter to incorporate them in new queries after close reading of a particular subset of newspaper articles. The key question derived from this method is: what do the word cloud results tell us? The results seem to indicate a meaningful connection between the concepts of race, wage and immigration in the Netherlands before the Second World War. The relatively large numbers of hits resulting from queries with combinations from these keywords are tempting clues to investigate this particular topic further. After all, it is obvious that the results from these queries alone do not demonstrate how these concepts were meaningfully connected; the researcher has to assess this connection by further alternating between query-guided distant reading and 'traditional' close reading of relevant texts (Huijnen et al., 2014: 79).

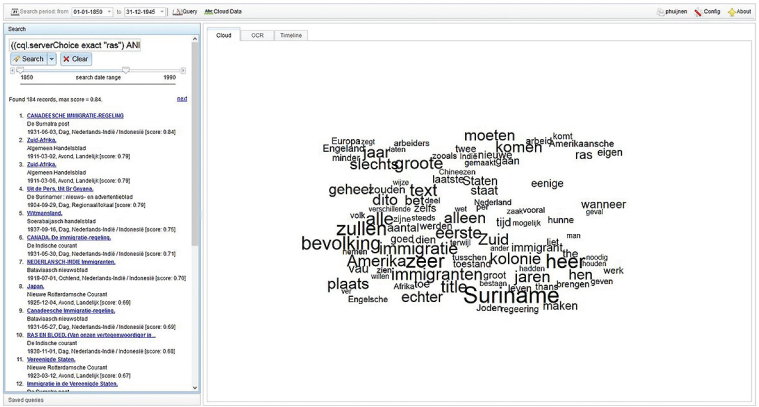

Figure 27.2: WAHSP search result plus word cloud based on the National Library of the Netherlands newspaper repository using the query 'race' AND 'immigration' ('ras' AND 'immigratie') for 1860–1945. (13 June 2013).

27.3 New Horizons: Historical Text Mining for Comparative Research as Part of The bilingual Text-Mining Tool Biland

An interdisciplinary team of researchers have tailored WAHSP to the language-specific needs of comparative historical research, with a particular focus on the identity, intensity, and location of discourses about heredity, genetics, and eugenics in Dutch and German newspapers between 1863 and 1940. The challenge has been to incorporate the semantics of the two languages (Dutch and German). A statistical machine translation service was included and used to translate existing lexicons and documents from Dutch and German (in both directions). xTAS's functionalities are used to leverage interactive creation, and the expansion and refinement of lexicons that are specific to the user's research questions and needs. xTAS feeds' visualisations allow users to examine the research domain along the aforementioned dimensions of time, context, and the identity and frequency of the discourse. As in WAHSP, BILAND employs a user-oriented, iterative model of collaboration between humanities scholars and ICT developers. Comparative, bilingual historical text mining raises a range of challenges. An important question is that of the linguistic comparability of the research topic as it is formulated in a specific query. The national vocabularies may not be literally translatable, for example, in the case of 'eugenics'. Whereas the Dutch terminology follows the English – *eugenetica*, *eugeniek* – in the German language the most common translation for eugenics is *Rassenhygiene* [racial hygiene]. The more literal translation *Eugenik* existed and was used in the same sense, but was not a sufficient keyword to explore eugenic notions in German newspapers. In this specific historical text mining, it is of utmost importance to be aware of the specifics of comparisons – a word or a concept, i.e., the idea behind that word (Huijnen et al., 2014: 81).

In addition to the comparability of historical concepts, the possibility of a comparison between given datasets should be tested. Do given datasets represent a similar historical entity – the public, the public debate (in an ideal situation)? In our media history case study the question is: is an equal range and coverage of newspapers represented in the dataset in a given period? Is there a comparable balance between national and regional newspapers, or newspapers representing urban and rural regions, etc.?

In BILAND, comparability is not yet possible. Because of IPR problems and the lack of useful digitised newspaper archives, the only digitised newspaper archive from Germany that this project was able to use was the *Amtspresse Preussens*. This dataset includes three 19th-century newspapers, with a total of less than 20,000 digitised pages.[10] These are hardly comparable to the Dutch dataset of 10 million pages, since the German dataset has neither the quantity, nor the wide time period or national scope of the Dutch dataset. However, German national libraries are rapidly catching up. They have initiated several digitising projects, e.g. within the Europeana community[11] or the *Deutsche Digitale Bibliothek*.[12]

Despite these IPR and digitisation challenges, the use of text-mining techniques holds promise as an innovative and exciting method for comparative international historical research. It can point to transnational concurrences or transfers of ideas, beliefs or knowledge in a far more time-efficient and validated way than traditional historical research has been able to do. Figure 27.3, for example, shows the concurrence of the word 'hygiene' in Dutch and German datasets.

[10] http://zefys.staatsbibliothek-berlin.de/ (accessed 03-02-2017)
[11] http://www.europeana.eu/portal/ (accessed 10-02-2016)
[12] https://www.deutsche-digitale-bibliothek.de/ (accessed 10-02-2016)

Figure 27.3: concurrence of the word 'hygiene' in Dutch and German datasets. Without ignoring the usual problems of historical comparison, the burst in 1863 in both sets of historical newspapers is significant enough to continue this line of research. (17 June 2013).

27.4 Up-Scaling WAHSP and BILAND for Larger Groups of Users: The Challenges

In 2013, the WAHSP and BILAND CLARIN-NL demonstrator projects were used as bases for two digital humanities projects in which historians and computer scientists cooperated. First, WAHSP modules were implemented and further developed in the open-source text-mining application Texcavator, which was developed as the supporting tool in the NWO-funded programme 'Translantis: Digital Humanities Approaches to Reference Cultures'.[13] One of the aims of the Translantis programme is to test and further develop Texcavator for historical research using 'big data' sets.[14] This allows a team of about ten historians to address conceptual historical questions on the role of reference cultures in 20th-century debates about social issues and collective identities, looking specifically at the emergence of the United States in public discourse in the Netherlands. The research team combines exploratory search and text mining to dig into the Delpher corpus of digitised historical newspapers and journals provided by the National Library of the Netherlands. Initially Texcavator combined xTAS with Elasticsearch as a natural follow-up of the WAHSP and BILAND tools.[15] However, from the beginning there were stability and scalability issues. In particular, the uptake of the tool was greater than foreseen, as were the underlying dynamically generated datasets, and the required high performance computing was underestimated. This resulted in a lack of responsiveness of the tool and user frustration. With the help of the eScience Center and SURFsara, and by eliminating the xTAS-supported tool functionalities, the stability and scalability were significantly improved. However, these performance gains came at the expense of the functionality profile of the tool.

The stable and scalable Texcavator tool is able to provide four main services: (1) allow researchers to carry out full-text searches and save queries, (2) produce normalised timelines showing how often (a combination of) keywords are produced in a specific period, (3) create word clouds

[13] http://translantis.wp.hum.uu.nl/ (accessed 03-02-2017)
[14] http://texcavator.hum.uu.nl/ (accessed 03-02-2017)
[15] https://github.com/elastic/elasticsearch (accessed 11-02-2016)

displaying the words used most often in the articles containing the entered keyword(s) (including the removal of stop words) and (4) export the query results to other text-analytics tools. The tool's main benefits are that it enables historians to trace specific words and changes in the contextualised word use over time and to keep systematic data logs. For named-entity recognition and sentiment mining (functionalities that were part of the original WAHSP tool), however, researchers now rely on other available open-source tools.

Second, the set of BILAND modules were implemented and further developed in the open-source multi-lingual text-mining application Asino, which is an integral part of the HERA-funded programme 'AsymEnc: Asymmetrical Encounters: Digital Humanities Approaches to Reference Cultures in Europe, 1815–1992'.[16] The AsymEnc Asino application was meant to be used as a multi-lingual text-mining tool to map the dynamics, intensity, and direction of intercultural references within European public discourse as represented in newspaper collections such as those of the British Library, the National Library of the Netherlands, and the Bibliothèque Nationale de Luxembourg. The AsymEnc research group developed this tool in order to trace and analyse regional, national, and European dimensions of cultural encounters, such as references to European urban centres, mass media, and consumer products. However, from the beginning, stability and scalability issues also became manifest, as had been the case for WAHSP. The computational expert group in Trier also chose Solr as an alternative to Elasticsearch and xTAS, and underestimated the time needed to integrate a machine translation service for Dutch, English, and German.[17] Moreover, the IPR rules for the cross-border exchange of digitised newspaper collections proved to be significantly stricter, due to commercial interests, than could have been foreseen at the beginning of the project. The historians involved have sought to overcome these limitations by using alternative open-source tools to perform named-entity recognition, topic modelling and GIS mapping of the locally available digitised newspaper collections, such as the *Pall Mall Gazette*,[18] while the Asino tool is being further developed by the Institute of Computer Science at the University of Göttingen (Coll Ardanuy et al., 2016).

27.5 Conclusion

Digital tools will enable historians to analyse massive volumes of texts and other big data sets and to integrate (socio-)linguistics, statistics, and geo-informatics in historical research. However, the technical and infrastructural requirements to meet those promises have not yet been fully realised. A crucial prerequisite for a productive digital research platform is the availability of high-performance computing facilities comparable with those used in physics and the life sciences in combination with a transdisciplinary working programme aimed at articulating and aligning the needs of the users. If these conditions are met, techniques of big data analytics will stimulate historians to set up new forms of systematic search trails that can, as we have shown, provide challenging perspectives on the circulation of ideas and notions in the public sphere. The evidence of a meaningful overlap and interference between drugs and eugenics as well as between wage and eugenics discourses certainly deserve further research.

Exploratory search methods that can provide a quick overview combined with tools to zoom into details on a predefined timeline are particularly useful for alternating effectively between distant reading and close reading. Our proposed integration of interactive exploratory search and text mining will support historians to set up systematic search trails. The tooling will help them interpret and contrast the returned result sets by exploring word associations for a result set, inspecting

[16] http://asymenc.wp.hum.uu.nl/ (accessed 11-02-2016)
[17] http://www.opensemanticsearch.org/ (accessed 11-02-2016)
[18] http://dhbenelux.org/wp-content/uploads/2015/04/42.pdf (accessed 11-02-2016)

the temporal distribution of documents and comparing selections so that more informed and principled document selection for close reading is possible. Obviously, this is no substitute for historical craftsmanship. WAHSP, BILAND, Texcavator and other open-source semantic text-mining tools are meant to be exploratory tools that ideally inspire new ideas and insights that would not have been generated through reading a small number of articles and can only be brought forth through the analysis of hundreds of articles. Insights gained by means of distant reading may help to frame new research questions, thus catalysing historical research. Digitally produced results often lead to unexpected associations that turn out to be promising for further research, but, in order to be meaningful, these require further conventional close reading.

There are a number of prerequisites that have to be met before digital tools can become standard procedure in historical research. First, it is essential that historians working with digital tools and building their arguments on digitally obtained research results be aware of what they are doing and keep meticulous lab logs of all their queries and visualisation results. This may sound obvious, but it does not always happen. Historians should have a clear understanding of the realities and causalities that word clouds, normalised time lines, or word vector approaches represent. They must also understand how to generate meaningful queries based on linguistic expertise. Historians should be able to interpret and explain text-mining research results in formulations such as, 'within the given digitized source material, in all articles containing word x and word y, word z also appears with a significant and normalized frequency'. This makes their arguments transparent and thus reproducible. Digital tools should not be treated as black boxes, with queries only going in and multiple visualisations mysteriously coming out and being used as evidence-based results. As Gibbs and Owens (2012) argue, '[t]he processes for working with the vast amounts of easily accessible and diverse large sets of data suggest a need for historians to formulate, articulate, and propagate ideas about how data should be approached in historical research'.

Demonstrator tools like WAHSP, BILAND and Texcavator offer explorative hints for certain lines of arguments, but do not automatically generate strong evidence or explanations for the arguments. The use of the aforementioned research tools indicates that in the public debate in the Netherlands at the start of the 20th century, drugs and inheritance were predominantly framed as medical, but the results of the text mining do not prove that this was true or explain why it was so.

In sum, open-source text-mining tools are not built to make historical scholarship obsolete, but rather to strengthen expertise through iteratively alternating distant and close reading, thus broadening heuristic capacities, and offering new analytical tools for data interpretation. These tools are meant to provide historians with new perspectives, and draw their attention through distant reading to potentially interesting cases that need further close reading. In this sense, it is evident that text mining can form a relevant addition to the historian's toolbox, and this outside the topics of drug and eugenic debates as well: text mining can be used to analyse cultural trends and patterns found in newspaper publications on a much broader scale.

Acknowledgements

This research was supported by the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement nr 288024 (LiMo- SINe project); by the Netherlands Organisation for Scientific Research (NWO) under project nrs 640.004.802, 727.011.005, 612.001.116, 317-52-010, HOR-11-10, Hor-11-19; by the Center for Creation, Content and Technology (CCCT), the WAHSP, BILAND and QuaMerdes projects funded by the CLARIN-nl program, the TROVe project funded by the CLARIAH program, Hera JRP Cultural Encounters 12-HERA-JRP-CE-FP-045 project, the Dutch national program COMMIT, the ESF Research Network Program ELIAS, the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences (KNAW), the Netherlands eScience Center under project number 027.012.105 and the Yahoo! Faculty.

References

Berry, D.M. (ed.): Understanding Digital Humanities. Palgrave Macmillan (2012)

Bingham, A.: The digitization of newspaper archives: Opportunities and challenges for historians. Twentieth Century British History 21(2), 225–231 (2010)

Burdick et al., A.: Digital Humanities. MIT Press (2012)

Coll Ardanuy M, Knauth, J. Beliankou A., Bos van den M., Sporleder C. Person-centric mining of historical newspaper collections. In. Volume 9819 of the series Lecture Notes in Computer Science (Springer, 2016), pp. 320–331.

Earheart, A.E., Jewell, A. (eds.): The American Literature Scholar in the Digital Age. University of Michigan Press (2011)

van Eijnatten, J., Pieters, T., Verheul, J.: Big data for global history: The transformative promise of digital humanities. Low Countries Historical Review 128, no.4, 55–77 (2013)

van Eijnatten J., Verheul J., Pieters T.: TS Tools: Using Texcavator to Map Public Discourse: TS: Tijdschrift voor Tijdschriftstudies 35, 59–65 (2014)

Gibbs, F., Owens, T.: The hermeneutics of data and historical writing. http://writinghistory.trincoll.edu/data/gibbs-owens-2012-spring/ (2012)

Graham, S., Milligan, I., Weingart, S.: The hermeneutics of data and historical writing. In: The Historian's Macroscope: Big Digital History. Imperial College Press (2013)

Hahn, D.: Modernisierung und Biopolitik: Sterilisation und Schwangerschaftsabbruch in Deutschland nach 1945. Campus (2000)

Huijnen P., Laan F., de Rijke M., Pieters T.: A digital humanities approach to the history of science; eugenics revisited in hidden debates by means of semantic text mining. In. A. Nadamoto et al (Eds): Soc Info 2013 Workshops, LNCS 8359 (Springer, New York), pp.71–85 (2014)

Huurnink, B., Hollink, L., van den Heuvel, W., de Rijke, M.: Search behavior of media professionals at an audiovisual archive: A transaction log analysis. Journal of the American Society for Information Science and Technology 61(6), 1180–1197 (June 2010)

Jackson, P., Moulinier, I.: Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization. John Benjamins, 2nd edn. (2007)

Jijkoun, V., de Rijke, M., Weerkamp, W.: Generating focused topic-specific sentiment lexicons. In: ACL '10 (2010)

Klausen, S., Bashford, A.: Fertility control: Eugenics, neo-malthusianism, and feminism. In: Bashford, A., Levine, S. (eds.) The Oxford Handbook of the History of Eugenics, pp. 98–115. Oxford University Press (2010)

Leonard, T.C.: Eugenics and economics in the progressive era. Journal of Economic Perspectives 19, 207–224 (2005)

Levine, P., Bashford, A.: Introduction: Eugenics and the modern world. In: Levine, P., Bashford, A. (eds.) The Oxford Handbook of the History of Eugenics, pp. 3–24. Oxford (2010)

Lombardo, P. (ed.): A Century of Eugenics in America: from the Indiana Experiment to the Human Genome Era. Indiana University Press (2001)

Meij, E., Bron, M., Huurnink, B., Hollink, L., de Rijke, M.: Learning semantic query suggestions. In: ISWC '09. Springer (2009)

Michel, J.B.: Quantitative analysis of culture using millions of digitized books. Science 6014, 176–183 (2010)

Nicholson, B.: The digital turn. Media History 19(1), 59–73 (2013)

Odijk D., de Rooij O., Peetz M-H., Pieters T., de Rijke M., Snelders S.: 'Semantic Document Selection', *TPDL 2012: Theory and Practice of Digital Libraries*: Springer. (2012)

Reulecke, J. (ed.): Herausforderung Bevölkerung: zu Entwicklungen des modernen Denkens über die Bevölkerung vor, im und nach dem 'Dritten Reich'. VS Verslag für Sozialwissenschaften (2007)

Seefeldt, D., Thomas III, W.G.: What is digital history? a look at some exemplar projects. Faculty Publications, Department of History. Paper 98. http://digitalcommons.unl.edu/historyfacpub/9 (2009)
Snelders, S., Pieters, T.: Van degeneratie tot individuele gezondheidsopties: Het maatschappelijk gebruik van erfelijkheidsconcepten in de twintigste eeuw. Gewina 26(4), 203–215 (2003)
Snelders S., Pieters T.: The blue lotus revisited: Public perceptions of drug use in the Dutch Empire, c. 1900-1942. Paper held at Drugs and drink in Asia: New perspectives from history. Conference of June 22–23, Shanghai. (2012)
Snelders S. Kaplan C. Pieters T.: On cannabis, chloral hydrate, and career cycles of psychotropic drugs in medicine. Bulletin of the History of Medicine 80, 95–114. (2006)
Turda, M.: Modernism and Eugenics. Palgrave MacMillan (2010)
Warwick, C., Terras, M.M., Nyhan, J.: Digital Humanities in Practice. Facet (2012)