TS•> Tijdschrift voor tijdschriftstudies
ISSN 1386-5870

Publisher: Igitur, Utrecht Publishing & Archiving Services. Website: www.tijdschriftstudies.nl

URN:NBN:NL:UI:10-1-114159. TS •> # 35, juli 2014, p. 59-65.

TS Tools: Using Texcavator to Map Public Discourse

ABSTRACT

In this article, Jaap Verheul, Toine Pieters, and Joris van Eijatten (Utrecht University) discuss the text mining tool Texcavator that they use in the Translantis research project, in order to map the emergence of the United States as a reference culture in the Netherlands between 1895 and 1995. Texcavator enables a fine grained analysis of large-scale document collections by using innovative text mining methodologies. It integrates a range of analytical tools such as concept clustering, sentiment mining, and Named Entity Recognition, to produce world clouds, time lines and other visualizations on-the-fly.

KEY WORDS

text mining, big data, periodicals, newspapers, reference cultures

TEXT MINING PUBLIC MEDIA

The text mining tool Texcavator has been developed for the research program Translantis: Digital Humanities Approaches to Reference Cultures; The Emergence of the United States in Public Discourse in the Netherlands, 1890-1990. This program is funded by the Netherlands Organization for Scientific Research (NWO) under the Horizon funding instrument, which focuses on innovation within the humanities and encourages research in wide-ranging, coherent programs that have the potential to determine the research agendas for future humanities research. The program started on January 1, 2013 and will run for five years. The scholarly aim of the program is to use digital technologies to analyze the role of reference cultures in debates about social issues and collective identities, looking specifically at the emergence of the United States in public discourse in the Netherlands from the end of the nineteenth century to the end of the Cold War. It introduces the concept of reference cultures to discuss long-term asymmetrical processes of cultural exchange involving dimensions of power and hegemony. This concept, which builds on recent academic debates in related fields such as transnational and global history, American studies, cultural transfer, and empire studies, recognizes the fact that some cultures assume a dominant role in the international circulation of knowledge and practices. They offer cultural references that others can adopt, adapt, or resist. These constructions of the 'other' can facilitate collective identity

formation. The conceptual framework of reference cultures will help us to understand how ideas, products, and practices associated with the United States were valued in Dutch public discourse in the long twentieth century.[1]

In order to understand how public discourse reflects and influences the emergence and impact of reference cultures, the Translantis program uses its semantic text mining tool Texcavator to analyze the largest repository of digitized historical periodicals that is currently available in the Netherlands, and is one of the largest national collections in the world, the digital newspaper archive of the National Library of the Netherlands (Koninklijke Bibliotheek). At present, this repository comprises over nine million pages from more than 200 different newspapers and periodicals published between 1618 and 1995, all together about 100 million articles. Not all national newspapers are represented in this collection – larger quality newspapers such as *de Volkskrant* and *NRC Handelsblad* are currently not included in the postwar period – and one could point at some other imbalances, for instance in the overrepresentation of newspapers published during the Second World War and in the overseas colonies, and the underrepresentation of recent decades. Nevertheless, in spite of such caveats, the available digitized newspapers offer an impressive sample of about 10 percent of all the newspapers that were published in the Netherlands. This is a solid basis for research on Dutch public periodicals.

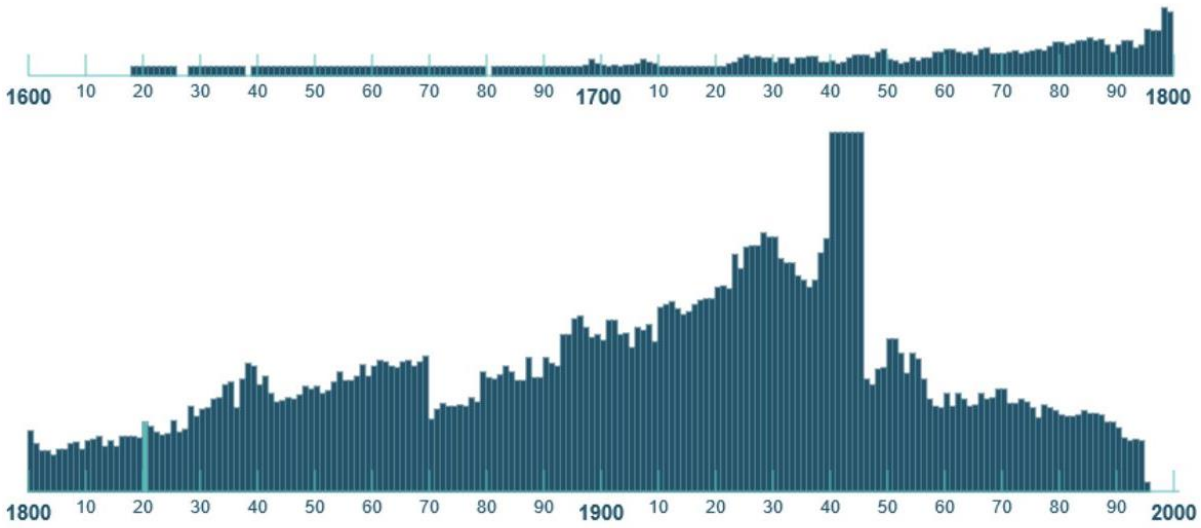

Fig. 1: Overview of digitized newspaper per year in the National Library of the Netherlands

[1] More about Translantis can be found on the project website. See for recent debates about cultural transfer and transnational history for instance P. Burke, *Cultural Hybridity*. Cambridge: Polity Press 2009; W. Fluck et al. (eds.), *Re-Framing the Transnational Turn in American Studies*. Hanover, N.H: Dartmouth College Press 2011; A. Iriye & P.-Y. Saunier, *The Palgrave Dictionary of Transnational History: From the Mid-19th Century to the Present Day*. Basingstoke: Palgrave Macmillan 2009; A. Iriye, *Global and Transnational History: The Past, Present, and Future*. Basingstoke: Palgrave Macmillan 2013.

The newspaper collection of the National Library is publicly available through its web interface Delpher, which also opens up about 1.5 million pages from 80 digitized journals published in the second half of the nineteenth century and the early twentieth century, and some 11 thousand historical books, mainly from the last decades of the eighteenth century.[2] For academic historians the research functionalities and options for storing search results in Delpher are limited. Texcavator, however, offers researchers an integrated set of analytical tools, visualization, and storage options that can be used by humanities researchers without specific computer skills.

Texcavator is based on a collection of open source text mining tools that have been developed by the team of Maarten de Rijke at the Intelligent Systems Lab at the University of Amsterdam (ISLA).[3] One of its precursors is the CLARIN-supported web application for historical sentiment mining in public media that is known under its acronym WAHSP. This tool was used by Toine Pieters and Stephen Snelders in a project that studied the historical dynamics of public debates on drugs, drug trafficking, and drug users in the early twentieth century (1900-1940). WAHSP not only helped to determine what terms were associated with drugs such as opium, morphine, heroin, and cocaine, but also offered quantitative evidence for the criminalization of the Dutch drug debates that took place around 1924. This tool was further developed in BILAND, a bilingual text mining tool that was developed to compare the identity, intensity, and location of discourses about heredity, genetics, and eugenics in Dutch and German newspapers between 1863 and 1940. Pim Huijnen and Toine Pieters employed the BILAND-tool to analyze to which extent eugenics debates in these two nations reflected social and cultural notions of individuals in relation to collective identities within the context of modernity. All these tools combined a set of text mining algorithms with a search engine that was able to search and index the textual data of the newspapers in the National Library in near real time. This provided researchers who are not trained in computer languages with unprecedented possibilities to access and analyze the big data collection of public media.[4]

---

[2] More about the Delpher newspapers collection of the National Library of the Netherlands can be found on kranten.delpher.nl.

[3] Most of these tools are based on the open source platform for text analytics xTAS. Examples are the Dutch Language Online Media Analysis (STEVIN), Building Rich Links to Enable Television History Research--BRIDGE (CATCH), Elite Network Shifts (KNAW), Infiniti (COMMIT), and Political Mashup. See xtas.net.

[4] An overview of these programs is offered by J. van Eijnatten et al., 'Big Data for Global History: The Transformative Promise of Digital Humanities.' *Low Countries Historical Review* 128:4, 2014, 55-77. See for examples of research results S. Snelders & T. Pieters, 'The Blue Lotus Revisited: Public Perceptions of Drug Use in the Dutch Empire, C. 1900-1942.' In *Drugs and Drink in Asia: New Perspectives from History: Proceedings of the Conference of June 22-23, 2012, Shanghai*, 2012; D. Odijk et al., 'Semantic Document Selection'. In P. Zaphiris et al. (eds.), *Theory and Practice of Digital Libraries*. Berlin: Springer 2012, 215-221; P. Huijnen et al., 'A Digital Humanities Approach to the History of Science.' In A. Nadamoto et al. (eds.), *Social Informatics*. Berlin: Springer 2014, 71-85.

FUNCTIONALITIES: ANALYTICS AND VISUALIZATIONS

Texcavator is a web based tool that combines a number of text mining analytical tools and visualizations that can be used to search historical newspapers. The web interface allows researchers to enter search terms. These can be individual words or word strings, but also complex queries that contain Boolean operators (such as AND, OR, and NOT) for a more focused search. The program is able to present the result of the query in various forms. It provides a list of records that match the search query, which contains the headline, title of the periodical, and date of each article, offering a direct link to the original in the National Library. A score indicates the relevance for the search string.

The program offers a number of visualization choices. One option is to generate a word cloud that is based on the total search result and represents the relative frequency of the selected words in the search result (figure 2). The program offers various possibilities to configure these clouds, for instance by adding stop words that can be eliminated from the search results (such as 'de', 'het' and 'en'). One of the most promising possibilities is to add the functionality of Named Entity Recognition to the word cloud differentiation option, so that entities such as proper names, dates, and geographical locations that are mentioned in the text are specifically represented in the word cloud in separate colors. Another powerful visualization tool makes it possible to select and indicate sentiment words in the cloud, enabling sentiment mining of the repository.

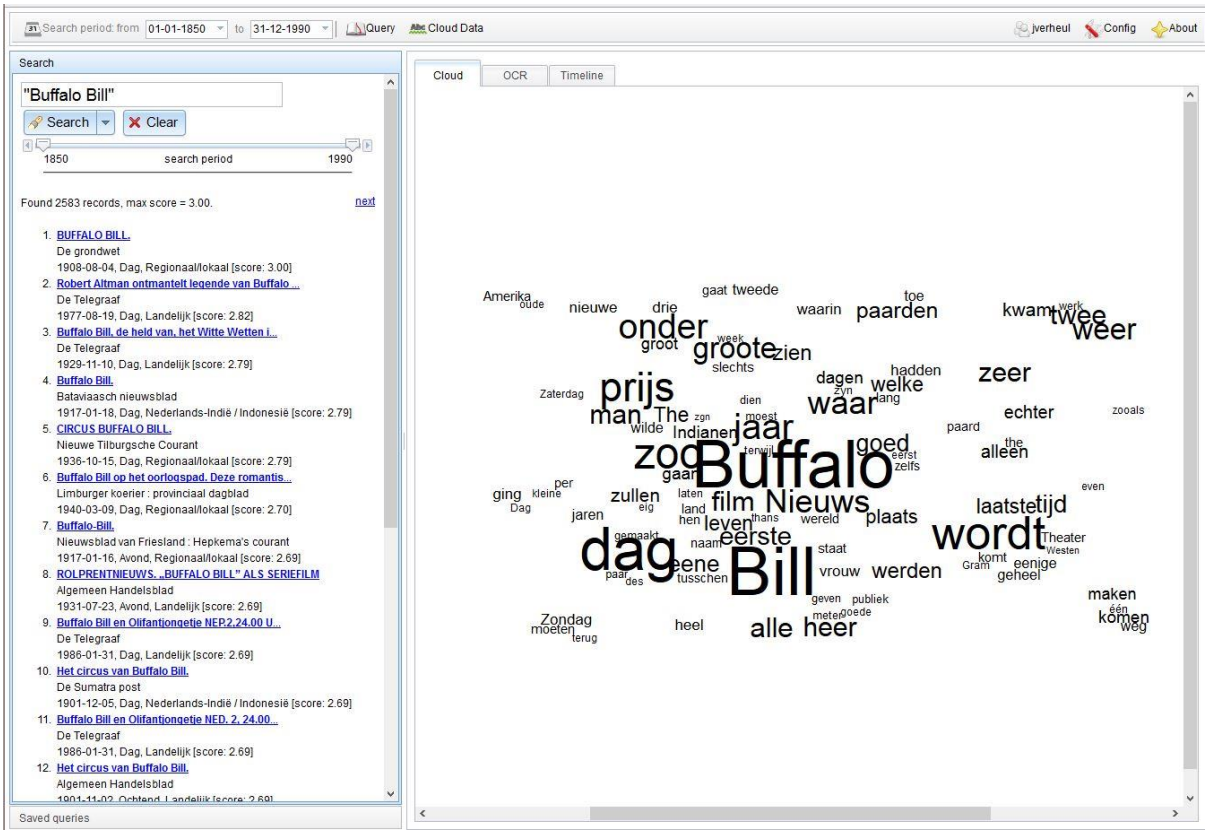

Fig. 2: Result of a query with a list of records and a word cloud

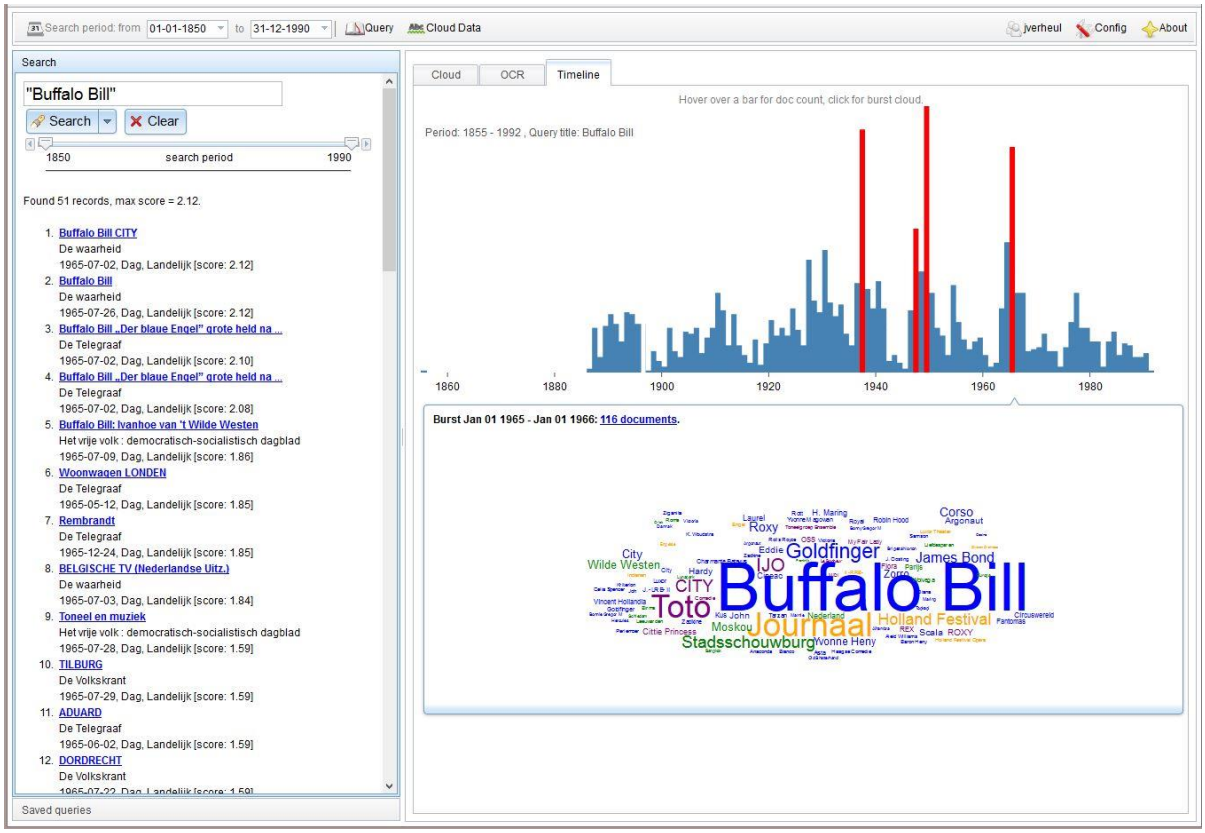

Fig. 3: Result of a query with a timeline and word cloud based on one 'burst' (indicated in red); the named entities in the word cloud are indicated by colors

Texcavator also offers the option of a historical timeline that is based on the search result (figure 3). This gives researchers a quick overview of the occurrences of searched terms over longer historical time periods. Years that yield a statistically higher score are indicated in red, and represent 'bursts'. Clicking on a burst yields a visualization of word associations of that burst alone and a list of documents contained within that burst. At any point in time the researcher can return to the original text of the newspaper article to obtain an in-depth understanding of what each burst is about. Together, these interactions allow historians to investigate the document selection interactively in order to detect patterns, improving the representativeness of the selection. The program can also present a scan of the selected newspaper article or the full text that is produced by Optical Character Recognition (figure 4). The quantitative data that are used to generate these visualizations, such as relative word frequencies counts, can be downloaded for later reference or post-processing.

CULTURAL TEXT MINING

The main added value of Texcavator lies in its possibilities for exploratory reading of historical patterns in public debates. In terms of methodology, semi-automatic document selection fits rather well with historical research as an alternative to manual browsing or random sampling, facilitating the combination of qualitative and quantitative approaches.

Through text mining and visualization, new insights can be gained from an initial selection. Word clouds depicting the linguistic context within which keywords occur are instrumental in helping historians with expert knowledge of the domain to combine and compare different historical periods in a free associative manner on the basis of a large number of historical documents. Exploring word associations and metadata, as well as visualizations of the documents over time, can lead to improved queries and, therefore, to a more representative document selection. Such quantitative analysis enhances the knowledge of historians.

Fig. 4: Texcavator: Result of a query with a scan of an individual article

The combination of close and distant reading is essential in this iterative research cycle. Close reading of the selected documents leads to a better understanding of the search results and will yield to new research questions and new or refined queries. Distant reading by means of word clouds and other kinds of histogram visualizations enables the researcher to cover a large set of data that would be impossible to scan manually. More importantly, this so-called bottom-up big data search strategy invariably leads to unexpected results that are not suggested by the academic literature. One of the most promising results from the WASHP and BILAND research projects is that distant reading can also point at 'hidden debates' in which important associated terms are not explicitly mentioned but are assumed by the participants in the debates. These missing indicators can be traced by the word clouds that the debates generate, which can point at shifts in associated meanings of key terms. Huijnen for instance traced implicit references to race

and unfit in debates about eugenics, which could be expected, but also in debates about social housing and minimum wage, where it was less obvious. In the Translantis project, Texcavator suggested geographical references associated with consumer products. For instance, sigarettes tended to be advertised with Oriental connotations in the early twentieth century, but gradually became associated with American life-style expressions in the course of the 1920s and 1930s, suggesting the emergence of the United States as a reference culture in Dutch public debates. These are promising possibilities for cultural text mining, which is mining of cultural aspects of entities and events. This is crucial in order to address macro-historical questions such as the emergence of transnational reference cultures.

Texcavator will become a tool for cultural text mining that enables scholars to search very large quantities of textual data simultaneously in a reliable and reproducible way. It enables the fine-grained analysis of large-scale document collections and offers state-of-the-art algorithms and visualizations that enable scholars to perform concept-clustering in big data sets and to distinguish long-term patterns in large news media repositories semi-automatically. Although Texcavator is currently not yet available to the wider research community as a result of copyright restrictions and limited server capacity, the aim is to produce a generic tool that can be used by other humanities scholars.

The Netherlands eScience Center has provided support to offer a scalable, stable, reliable IT-environment for Texcavator. This support will significantly strengthen the employment of computational methods in this project by guaranteeing scalability and stability and will point out new directions for the humanities.

•> JORIS VAN EIJNATTEN, TOINE PIETERS, *and* JAAP VERHEUL *are applicants and project leaders of the NWO-funded Translantis research project.*