Paper

Eugenics revisited in hidden debates by means of multilingual semantic text-mining with the Biland demonstrator

Toine Pieters[1] , Pim Huijnen[1], Maarten de Rijke[2]

[1] Descartes Centre for the History and Philosophy of the Sciences and the Arts, Utrecht University, the Netherlands
t.pieters@uu.nl, p.huijnen@uu.nl,
2 Intelligent Systems Lab Amsterdam, University of Amsterdam, the Netherlands
derijke@uva.nl

Introduction

The Biland Clarin demonstrator has been tailored to the language-specific needs of comparative historical research, with a specific focus on the identity, intensity, and location of discourses about heredity, genetics, and eugenics in Dutch and German newspapers between 1865 and 1900. The challenge has been to incorporate the semantics of two different languages (in this case Dutch and German) and scripts (such as Latin and Gothic). The BILAND project employed a user-oriented, iterative model of collaboration between humanities scholars and ICT developers. Every developmental task and research activity envisaged within the project is a transdisciplinary co-production. This included selecting and filtering out meaningful lexical items, carrying out text-mining tasks, training the algorithms, and meeting the needs of the domain users by realizing feed-back loops. The goal was to analyse to what extent eugenics debates in the Netherlands and Germany reflected social and cultural notions of individual in relation to collective identities within the context of modernity. We started to focus on the multiple discourses that converged around the use and adaptation of genetic knowledge and eugenics in the workplace, the home and the wider world. The challenge was to qualify and quantify these 'hidden debates'. I will show how Biland enabled us not only to mine the obvious heredity and eugenics related terms but also explore the more unconscious, latent use of heredity or eugenics related notions and ideas in newspapers.

Available big data repositories

The Dutch repository that has been available for text-mining is the newspaper archive of the National Library of the Netherlands (*Koninklijke Bibliotheek*). At present, this repository comprises over 10 million pages from more than 200 different newspapers and periodicals published between 1618 and 1995, all together about 100 million articles.[1] The available German repository was relatively limited in size. Because of IPR problems and the lack of useful digitized newspaper archives, the only digitized newspaper archive, the only digitized newspaper archive from Germany the project was permitted to make use of was the so-called Amtspresse Preussens that was digitized in a pilot project of the Staatsbibliothek zu Berlin.[2] The Amtspresse Preussens dataset comprises of three 19th century newspapers[3], together containing less than 20,000 digitized pages in the period 1860-1900. These are hardly comparable to the Dutch data set of 10 million pages, not only in quantity, but also in the time period covered and the national scope. The German national libraries are, however, rapidly catching up. They have initiated several digitizing projects, among others within the Europeana[4] community. Despite the quantitative differences we were able to use comparable data test-sets for multi-lingual text mining text-mining.

Text mining tool specifications

The technical basis of Biland is an ElasticSearch instance combined with the xTAS text analysis service. xTAS includes modules for online and offline processing. xTAS provides other essential text pre-processing modules (morphologically normalization, format and encoding reconciliation, named entity recognition and normalization, etc). xTAS has been developed by the Intelligent Systems Lab at the University of Amsterdam (ISLA).[1] This open source platform for text analytics has also been applied and tested in

[1] http://kranten.delpher.nl/
[2] http:// zefys.staatsbibliothek-berlin.de/amtspresse.
[3] Provinzial Correspondenz (1863-1884), Neueste Mittheilungen (1882-1894) and Teltower Kreisblatt (1856-1896).
[4] http://www.europeana.eu/

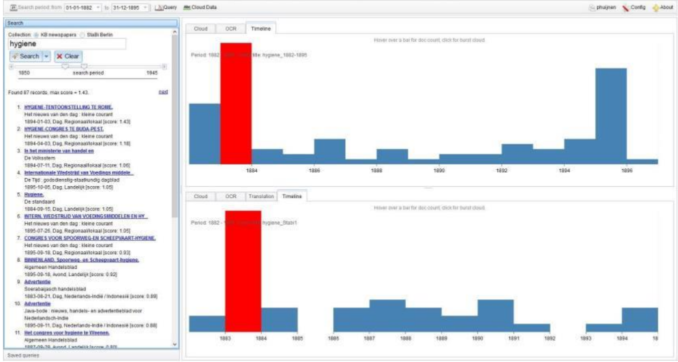
computational humanities projects such as Dutch Language Online Media Analysis STEVIN), Building Rich Links to Enable Television History Research--BRIDGE (CATCH), Elite Network Shifts (KNAW), Infiniti (COMMIT), and Political Mashup.[2] Biland comes with visualization modules built in D3.js (interactive wordclouds and timelines). A statistical machine translation service is also available, which can be used to translate existing lexicons and documents between Dutch and German (both directions). The functionalities of xTAS are used to leverage interactive creation, expansion and refinement of lexicon's specific to the user's research questions and needs. xTas feeds visualizations that allow users to examine the research domain along the aforementioned dimensions of time, context, and the identity and frequency of the discourse.

Results

The use of multilingual text mining techniques holds the promise of an innovative and exciting method for comparative historical research. In principle, new digital tools like Biland are is able to address the history of concepts and of mentalities in creative new ways. It can point at concurrences or transfers of ideas, beliefs or knowledge that traditional historical research is not able to do. Figure 1, for example, shows the concurrence of the word 'hygiene'[5] in both Dutch and German datasets. Without ignoring the usual problems of historical comparison, the burst in 1863 in both sets of historical newspapers is exciting enough to continue this line of research.

Conclusion

Digital tools offer historians revolutionary research opportunities to analyze massive volumes of texts and other big data sets and to integrate (socio-) linguistics, statistics and geo-informatics into historical research. Our proposed combination of interactive exploratory search and text mining supports historians to set up systematic search trails; the tooling helps them interpret and contrast the returned multilingual result sets. By exploring word associations for a result set, inspecting the temporal distribution of documents and by comparing selections historians can combine new forms of close and distant reading. Obviously, this is no substitute for the historical workmanship. Rather, BILAND is meant as a heuristic tool that ideally brings the historian new insights that help to frame new research questions, thus catalysing the research process (1).



Search period from 01-01-1882 to 31-12-1895

Collection: ● KB newspapers ○ Stabi Berlin

hygiene

Found 87 records; max score = 1.43.

1. HYGIENE-TENTOONSTELLING TE ROME.
Het nieuws van den dag : kleine courant
1894-01-03, Dag, Regionaal/lokaal [score: 1.43]
2. HYGIENE-CONGRES TE BUDA-PEST.
Het nieuws van den dag : kleine courant
1894-04-03, Dag, Regionaal/lokaal [score: 1.18]
3. In het ministerie van handel en
De Volksstem
1894-07-11, Dag, Regionaal/lokaal [score: 1.06]
4. Internationale Wedstrijd van Voedings middele...
De Tijd : godsdienstig-staatkundig dagblad
1895-10-05, Dag, Landelijk [score: 1.05]
5. Hygiene.
De standaard
1884-09-15, Dag, Landelijk [score: 1.05]
6. INTERN. WEDSTRIJD VAN VOEDINGSMIDDELEN EN HY...
Het nieuws van den dag : kleine courant
1895-07-26, Dag, Regionaal/lokaal [score: 1.05]
7. CONGRES VOOR SPOORWEG-EN SCHEEPVAART-HYGIENE.
Het nieuws van den dag : kleine courant
1895-09-18, Dag, Regionaal/lokaal [score: 0.93]
8. BINNENLAND. Spoorweg- en Scheepvaart-hygiene.
Algemeen Handelsblad
1895-09-18, Avond, Landelijk [score: 0.92]
9. Advertentie
Soerabaijasch handelsblad
1883-08-21, Dag, Nederlands-Indie / Indonesië [score: 0.89]
10. Advertentie
Java-bode : nieuws, handels- en advertentieblad voor Nederlandsch-Indie
1895-09-11, Dag, Nederlands-Indie / Indonesië [score: 0.88]
11. Het congres voor hygiene te Weenen.
Algemeen Handelsblad

Period: 1882 - ... title: hygiene_1882-1895

Period: 1882 - ... hygiene_Stabi1

Figure 1

BILAND search result in the form of two timelines from the query 'hygiene' for 1860-1900. The Dutch timeline is shown on top, the German below

(1) 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. Natamoto et al (eds); Socinfo 2013 Workshops, 1. NCS 8359, pp. 71-85, (Springer, New York, 2014).

[1]http://xTAS.science.uva.nl.

[2] http://ilps.science.uva.nl/biblio/duoman-dutch-language-online-media-analysis; http://ilps.science.uva.nl/node/735; http://ilps.science.uva.nl/news/knaw-computational-humanities-grant; http://ilps.science.uva.nl/research/projects/bridge; http://www.kitlv.nl/home/Projects?id=25; http://www.project-infiniti.nl/; http://politicalmashup.nl/.

[5] In Dutch: 'hygiëne', in German: 'Hygiene'.