Mining the Twentieth Century's History from the Time Magazine Corpus

Kestemont, Mike; Karsdorp, F.B.; Düring, Marten

published in
Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)
2014

document version
Publisher's PDF, also known as Version of record

Mining the Twentieth Century's History from the Time Magazine Corpus

Mike Kestemont
University of Antwerp
Prinsstraat 13, D.188
B-2000, Antwerp
Belgium
mike.kestemont
@uantwerpen.be

Folgert Karsdorp
Meertens Institute
Postbus 94264
1090 GG Amsterdam
The Netherlands
Folgert.Karsdorp
@meertens.knaw.nl

Marten Düring
University of North-Carolina
551 Hamilton Hall
CB 3195, Chapel Hill
North Carolina 27599
United States
marten@live.unc.edu

Abstract

In this paper we report on an explorative study of the history of the twentieth century from a lexical point of view. As data, we use a diachronic collection of 270,000+ English-language articles harvested from the electronic archive of the well-known *Time Magazine* (1923–2006). We attempt to automatically identify significant shifts in the vocabulary used in this corpus using efficient, yet unsupervised computational methods, such as Parsimonious Language Models. We offer a qualitative interpretation of the outcome of our experiments in the light of momentous events in the twentieth century, such as the Second World War or the rise of the Internet. This paper follows up on a recent string of frequentist approaches to studying cultural history ('Culturomics'), in which the evolution of human culture is studied from a quantitative perspective, on the basis of lexical statistics extracted from large, textual data sets.

1 Introduction: Culturomics

Although traditionally, the Humanities have been more strongly associated with qualitative rather than quantitative methodologies, it is hard to miss that 'hipster' terms like 'Computational Analysis', 'Big Data' and 'Digitisation', are currently trending in Humanities scholarship. In the international initiative of Digital Humanities, researchers from various disciplines are increasingly exploring novel, computational means to interact with their object of research. Often, this is done in collaboration with researchers from Computational Linguistics, who seem to have adopted quantitative approaches relatively sooner than other Humanities disciplines. The subfield of Digital History (Zaagsma, 2013), in which the present paper is to be situated, is but one of the multiple Humanities disciplines in which rapid progress is being made as to the application of computational methods. Although the vibrant domain of Digital History cannot be exhaustively surveyed here due to space limits, it is nevertheless interesting to refer to a recent string of frequentist lexical approaches to the study of human history, and the evolution of human culture in particular: 'Culturomics'.

This line of computational, typically data-intensive research seeks to study various aspects of human history, by researching the ways in which (predominantly cultural) phenomena are reflected in, for instance, word frequency statistics extracted from large textual data sets. The field has been initiated in a lively, yet controversial publication by Michel et al. (2011), which – while it has invited a lot of attention in popular media – has not gone uncriticized in the international community of Humanities.[1] In this paper, the authors show how a number of major historical events show interesting correlations with word counts in a vast corpus of *n*-grams extracted from the Google Books, allegedly containing 4% percent of all books ever printed.

In recent years, the term 'Culturomics' seems to have become an umbrella term for studies engaging, often at an increasing level of complexity, with the seminal, publicly available Google Books NGram Corpus (Juola, 2013; Twenge et al., 2012; Acerbi et al., 2013b). Other studies, like the inspiring contribution by Leetaru (2011) have independently explored other data sets for similar purposes, such as the retroactive prediction of the Arab Spring Revolution using news data. In

[1] Consult, for instance, the critical report by A. Grafton on the occasion of a presentation by Michel and Lieberman Aiden at one of the annual meetings of the American Historical Association (https://www.historians.org/publications-and-directories/perspectives-on-history/march-2011/loneliness-and-freedom).

Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014, pages 62–70,
Gothenburg, Sweden, April 26 2014.

the present paper, we seek to join this recent line of Culturomics research: we will discuss a series of quantitative explorations of the *Time Magazine Corpus* (1923–2006), a balanced textual data set covering a good deal of the twentieth (and early twenty-first) century.

The structure of the paper is as follows: in the following section 2, we will discuss the data set used. In section 3, we will introduce some of the fundamental assumptions underlying the Culturomics approach for Big Data, and report on an experiment that replicates an earlier sentiment-related analysis of the Google Books Corpus (Acerbi et al., 2013b) using our *Time* data. Subsequently, we will apply a Parsimonious Language Model to our data (section 4) and assess from a qualitative perspective how and whether this technique can be used to extract the characteristic vocabulary from specific time periods. To conclude, we will use these Parsimonious Language Models in a variability-based neighbor clustering (section 5), in an explorative attempt to computationally identify major turning points in the twentieth century's history.

2 Data: Time Magazine Corpus

For the present research, we have used a collection of electronic articles harvested from the archive of the well-known weekly publication *Time Magazine*. The magazine's online archive is protected by international copyright law and it can only be consulted via a paying subscription.[2] Therefore, the corpus cannot be freely redistributed in any format. To construct the corpus, we have used metadata provided by corpus linguist Mark Davies who has published a searchable interface to the Time Corpus (Davies, 2013). For the present paper, we were only dependent on the unique identification number and publication year which Davies provides for each article. Users who are interested in downloading (a portion of) the corpus which we used, can use this metadata to replicate our findings.

We have used the Stanford CoreNLP Suite to annotate this collection (with its default settings for the English language).[3] We have tokenized and lemmatized the corpus with this tool suite. Additionally, we have applied part-of-speech tagging (Toutanova et al., 2003) and named entity recognition (Finkel et al., 2005). In the end, our reconstructed version of the Time Corpus in total amounted to 270,695 individual articles. In its entirety, the corpus counted 191,743,541 distinct word forms (including punctuation marks), 867,399 forms of which proved unique in their lowercased format. Some general statistics about our reconstructed version of the Time Corpus are given in Table 1. In addition to the cumulative word count statistics about the corpus, we have included the frequency information per decade (1920s, 1930s, etc.), as this periodisation will prove important for the experiments described in section 4.

Period	# Documents	# Word forms	# Unique forms
1920s	24,332	11,155,681	158,443
1930s	32,788	20,622,526	222,777
1940s	41,832	22,547,958	234,918
1950s	42,249	25,638,032	251,658
1960s	35,440	27,355,389	258,276
1970s	27,804	25,449,488	218,322
1980s	25,651	24,185,889	208,678
1990s	23,300	20,637,179	204,393
2000s	17,299	14,151,399	176,515
Overall	**270,695**	**191,743,541**	**867,399**

Table 1: General word frequency statistics on the reconstructed version of the Time Corpus (1923-2006).

In its entirety, the corpus covers the period March 1923 throughout December 2006. It only includes articles from the so-called 'U.S. edition' of *Time* (i.e., it does not contain articles which only featured in the e.g. European edition of the Magazine). Because of *Time*'s remarkably continuous publication history, as well as the considerable attention the magazine traditionally pays to international affairs and politics, the Time Corpus can be expected to offer an interesting, albeit exclusively American perspective on the recent world history. As far as we know, the corpus has only been used so far in corpus linguistic publications and we do not know of any advanced studies in the field of cultural history that make extensive use of the corpus.

3 Assumption: Lexical frequency

Previous contributions to the field of Culturomics all have in common that they attempt to establish a correlation between word frequency statistics and cultural phenomena. While this is rarely explicitly voiced, the broader assumption underlying these studies is that frequency statistics extracted from

[2] http://content.time.com/time/archive.

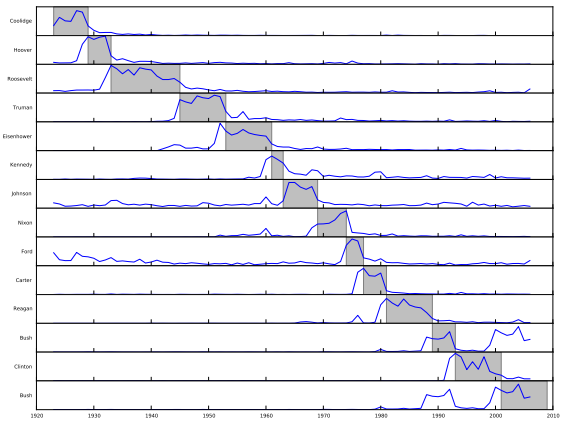
[3] http://nlp.stanford.edu/software/corenlp.shtml

the texts produced by a society at specific moment in history, will necessarily reflect that society's cultural specific (e.g. cultural) concerns at that time. As such, it can for instance be expected that the frequency of conflict-related terminology will tend to be more elevated in texts produced by a society at war than one at peace. (Needless to say, this need not imply that a society e.g. supports that war, since the same conflict-related terminology will be frequent in texts that oppose a particular conflict.) Obviously, the resulting assumption is that the study of developments in the vocabulary of a large body of texts should enable the study of the evolution of the broader historical concerns that exist(ed) in the culture in which these texts were produced.

Frequency has been considered a key measure in recent studies into cultural influence (Skiena and Ward, 2013). The more frequent a word in a corpus, the more weighty the cultural concerns which that word might be related to. A naive illustration of this frequency effect can be gleaned from Figure 1. In the subplots of the figure, we have plotted the absolute frequency with which the last names of U.S. presidents have been yearly mentioned throughout the Time Corpus (in their lowercased form, and only when tagged as a named entity). The horizontal axis represents time, with grey zones indicating start and end dates of the administration periods. The absolute frequencies have been normalised in each year, by taking their ratio over the frequency of the definite article *the*. Before plotting, these relative frequencies have been mean-normalised. (Readers are kindly requested to zoom in on the digital PDF to view more detail for all figures.) Although this is by no means a life-changing observation, each presidential reign is indeed clearly characterised by a significant boost in the frequency of the corresponding president's last name. Nevertheless, the graph also displays some notable deficiencies, such the confusion of father and son Bush, or the increase in frequency right before an administration period, which seems related to the presidential election campaigns.

Importantly, it has been stressed that reliable frequency information can only be extracted from large enough corpora, in order to escape the bias caused by limiting oneself to e.g. too restricted a number of topics or text varieties. This has caused studies to stress the importance of so-called 'Big

Figure 1: Diachronic visualisation of mean-normalised frequencies-of-mention of the last names of U.S. presidents in the Time corpus, together with their administration periods.



Data' when it comes to Culturomics, reviving the old adagium from the field of Machine Learning 'There's no data like more data', attributed to Mercer. In terms of data size, it is therefore an important question whether the Time Corpus is a reliable enough resource for practicing Culturomics. While the Time Corpus (just under 200 million tokens) is not a small data set, it is of course orders of magnitude smaller than the Google Books corpus with its intimidating 361 billion words. As such, the Time Corpus might hardly qualify as 'Big Data' in the eyes of many contemporary data scientists. One distinct advantage which the Time Corpus might offer to counter-balance the disadvantage of its limited size, is the high quality, both of the actual text, as well as the metadata (OCR-errors are for instance extremely rare).

In order to assess whether a smaller, yet higher-quality corpus like the Time Corpus might yield valid results when it comes to Culturomics we have attempted to replicate an interesting experiment reported by Acerbi et al. (2013b) in the context of a paper on the expression of emotions in twentieth century books. For their research, they used the publicly available Google Books unigram corpus. In our Figure 2 we have reproduced their 'Figure 1: Historical periods of positive and negative moods'. For this analysis, they used the so-called LIWC-procedure: a methodology which attempts to measure the presence of particular emotions in texts by calculating the relative occurrences of a set of key words (Tausczik and Pen-

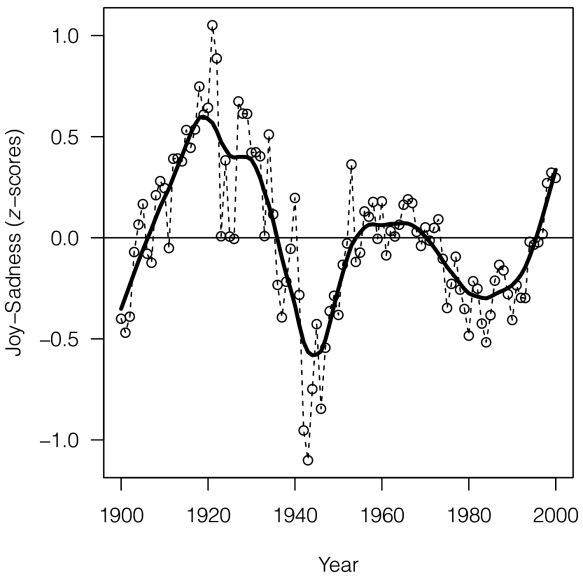
nebaker, 2010).[4] In the authors' own words, the graph "shows that moods tracked broad historical trends, including a 'sad' peak corresponding to Second World War, and two 'happy' peaks, one in the 1920's and the other in the 1960's."

We have exactly re-engineered their methodology and applied it to the Time Corpus. The result of this entirely parallel LIWC-analysis (Tausczik and Pennebaker, 2010) of the Time Corpus is visualized in Figure 3. While our data of course only starts in 1923 instead of 1900 (cf. grey area), it is clear that our experiment has produced a surprisingly similar curve, especially when it comes to the 'sad' and 'happy' periods in the 1940s and 1960s respectively. These pronounced similarities are especially remarkable because, to our knowledge, the Time Corpus is not only much smaller but also completely unrelated to the Google Books corpus. This experiment thus serves to emphasise the remarkable stability of certain cultural trends as reflected across various text types and unrelated text corpora.[5] Moreover, these results suggest that the Time Corpus, in spite of limited size, might still yield interesting and valid results in the context of Culturomics research.

4 Parsimonious Language Models

As discussed above, Michel et al. (2011) have proposed a methodology in their seminal paper, whereby, broadly speaking, they try to establish a correlation between historical events and word counts in corpora. They show, for instance, that the term 'Great War' is only frequent in their data until the 1940s: at that point the more distinctive terms 'World War I' and 'World War II' suddenly become more frequent. One interesting issue here is that this methodology is characterised by a modest form a 'cherry picking': with this way of working, a researcher will only try out word frequency plots of which (s)he expects beforehand that they will display interesting trends. Inevitably, this fairly supervised approach might lower one's chance to discover new phenomena, and thus reduces the chance for scientific serendipity to occur. An equally interesting, yet much less supervised approach might therefore be to *automatically* identify which terms are characteristic for a given time span in a corpus.

Figure 2: Figure reproduced from Acerbi et al. (2013b): 'Figure 1: Historical periods of positive and negative moods'. Also see Figure 3.

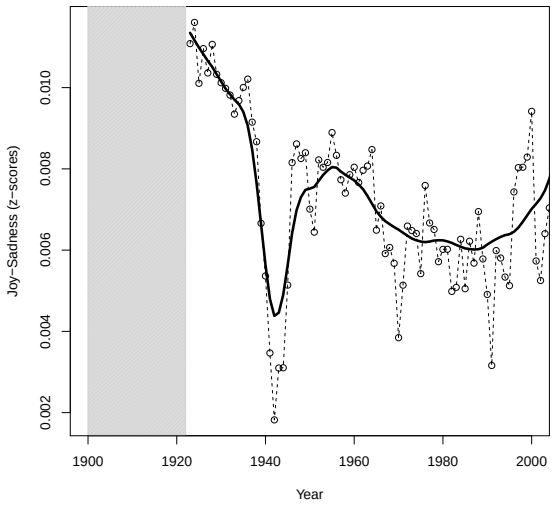


In this respect, it is interesting to refer to Parsimonious Language Models (PLMs), a fairly recent addition to the field of Information Retrieval (Hiemstra et al., 2004). PLMs can be used to create a probabilistic model of a text collection, describing the relevance of words in individual documents in contrast to all other texts in the collection. From the point of view of indexing in Information Retrieval, the question which a PLM in reality tries to answer is: 'Suppose that in the future, a user will be looking for this document, which search terms is (s)he most likely to use?' As such, PLMs offers a powerful alternative to the established TF-IDF metric, in that they are also able to estimate which words are most characteristic of a given document. While PLMs are completely unsupervised (i.e. no manual annotation of documents is needed), they do require setting the λ parameter beforehand. The λ parameter will of course have a major influence on the final results, since it will control the rate at which the language of each document will grow different from that of all other documents, during the subsequent updates of the model. (For the mathematical details on λ, consult Hiemstra et al. (2004).) Thus, PLMs can be expected to single out more characteristic vocab-

[4] We would like to thank Ben Verhoeven for sharing his LIWC-implementation. The methodology adopted by Acerbi et al. (2013b) has been detailed in the following blog post: `http://acerbialberto.wordpress.com/tag/emotion/`.

[5] Acerbi et al. (2013a) have studied the robustness of their own experiments recently, using different metrics.

Figure 3: LIWC-analysis carried on the Time Corpus (cf. Figure 2), attempting to replicate the trends found by Acerbi et al. (2013b). Plotted is the absolute difference between the z-scores for the LIWC-categories 'Positive emotions' and 'Negative emotions'. The same smoother ('Friedman's supersmoother') has been applied (R Core Team, 2013).



ulary than simpler frequentist approaches and, interestingly, they are more lightweight to run than e.g. temporal topic models.

In a series of explorative experiments, we have applied PLMs to the Time Corpus. In particular, we have build PLMs for this data, by combining individual articles into much larger documents: both for each year in the data, as well as all 'decades' (e.g. $1930 = 1930-1939$) we have constructed such large, multi-article documents. For both document types (years and decades), we have subsequently generated PLMs. In Figure 4 to Figure 12 we have plotted the results for the PLMs based on the decade documents (for $\lambda = 0.1$). In the left subpanel, we show the 25 words (technically, the lowercased lemma's) which the PLM estimated to be most discriminative for a given decade. In the right subpanel, we have plotted the evolution of the relevance scores for each of these 25 words in the year-based PLM (the grey zone indicates the decade). Higher scores indicate a more pronounced relevance for a given decade. For the sake of interpretation, we have restricted our analysis to words which were tagged as nouns ('NN' & 'NNS').

Figure 4: PLM for the 1920s.

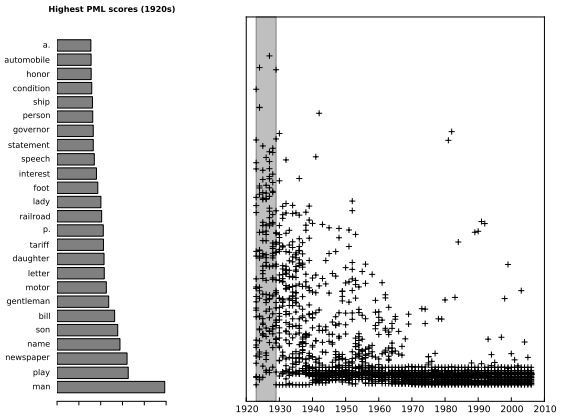
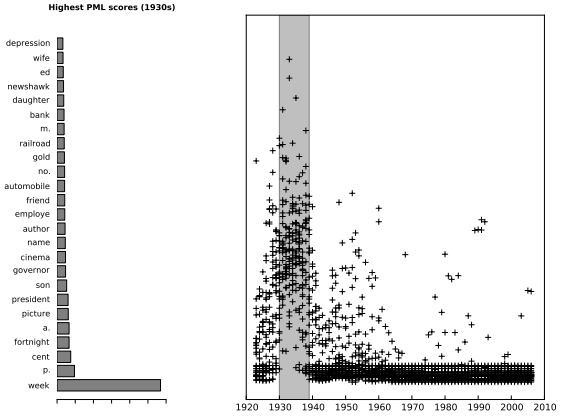


Figure 5: PLM for the 1930s.



It is not immediately clear how the output of these PLMs can be evaluated using quantitative means. A concise qualitative discussion seems most appropriate to assess the results. For this reason, we have combined individual articles into larger decade documents in these experiments, since this offers a very intuitive manner of arranging the available sources from a historical, interpretative point of view. Often, when people address the periodisation of the twentieth century they will use decades, where terms like e.g. 'the seventies' or 'the twenties' refer to a fairly well-delineated concept in people's minds, associated with a particular set of political events, people and cultural phenomena, etc. By sticking to this decade-based periodisation, we can verify fairly easily to what extent the top 25 yielded by the PLM corresponds to commonplace historical

Figure 6: PLM for the 1940s.

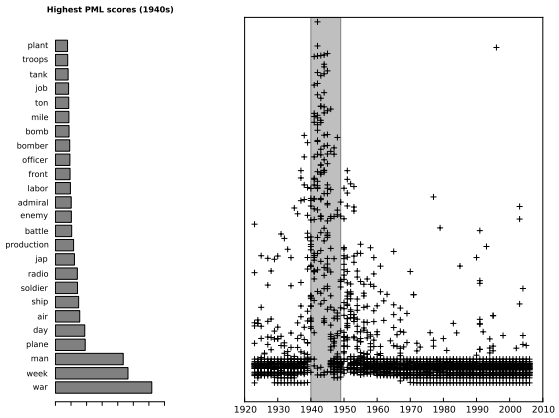


Figure 7: PLM for the 1950s.

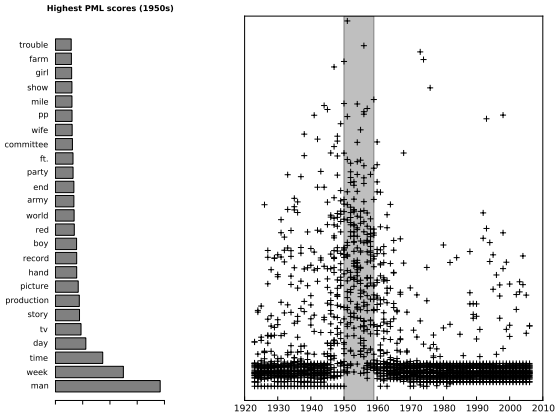


Figure 8: PLM for the 1960s.

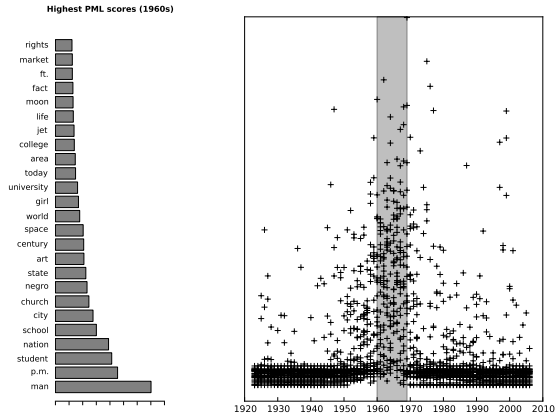


Figure 9: PLM for the 1970s.

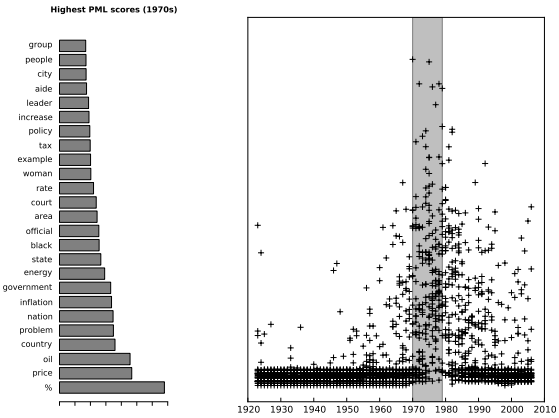


Figure 10: PLM for the 1980s.

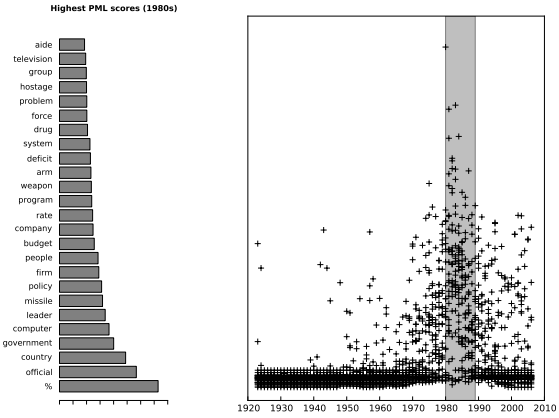


Figure 11: PLM for the 1990s.

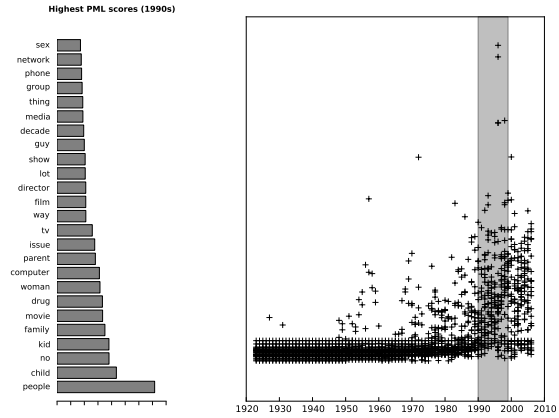
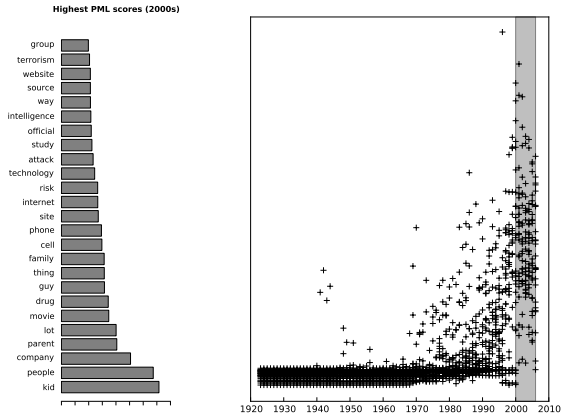


Figure 12: PLM for the 2000s.



stereotypes about the decades in the twentieth century.

Let us start by inspecting the top 25 for the 1940s, a decade in which the Second War II naturally played a major role. Already at first glance, it is clear that the top 25 is dominated by war-related terminology (*war*, *soldier*, *enemy*, ...). Interestingly, the list also contains words referring to WWII, but not from the politically correct jargon which we would nowadays use to address the issue (e.g. *jap*). Remarkable is the pronounced position of aviary vocabulary (*bomber*, *air*, *plane*, ...), which is perhaps less surprising if we consider the fact that WWI was one of the first international conflicts in which aircrafts played a major military role.

Interestingly, the 1920s are hardly characterised by an equally focused set of relevant words. Although mobility does seem to play an important role (cf. the recently invented *automobile*, but also *ship* and *railroad*), a number of less meaningful abbreviations (such as *p.* for 'page') pop up that seem connected to superficial changes in *Time*'s editorial policies, rather than cultural developments. (Future analyses might want to remove such words manually.) On the other hand, the use of the terms *lady* and *honour* might be rooted in a cultural climate that is different from ours ('lady' seems the equivalent of woman today). A number of parallel observations can be made for the 1930s, although here, the high ranking word *depression* is of course striking (cf. the economic crisis of 1929). Fascinatingly, a variety of denominations for (popular) media play a major role throughout the decade PLMs. Note, that while the 1920s' top 25 mentioned the *radio* as the primary communication medium, the popular *cinema* and (moving?) *picture* show up in the 1930s. Interestingly, the popular media of *tv* and *record* make their appearance in the 1950s. (In the 1980s and 1990s top 25, *television* moreover continues to show up.)

The PLM also seems to offer an excellent cultural characterisation of the 1960s and the associated baby boom, with an emphasis on the controversies of the time, debate involving human rights (*rights*, *negro*, *nation*), and in particular educational (*college*, *university*, *school*). The use of 'educational' words might well be related to the social unrest, much of which took place in and around universities. Does the striking presence of the word 'today' in the list reveal an elevated *hic et nunc* mentality in the contemporary States? America's well-documented interest in space traveling at the time is also appropriately reflected (*space*, *moon*). Perhaps unexpectedly, this seemingly optimistic 'Zeitgeist' is more strongly associated in the Time Corpus with the sixties, than with the seventies: in the flower-power era, *Time* displays a remarkable focus on political and especially economic issues. Rather, the oil crisis seems to dominate Time's lexis in the seventies.

In the 1990s and 2000s, we can observe a focus on what one might unrespectfully call 'first-world problems', involving for instance family relations (*family*, *kid*, *parent*, *child*, etc.). Apart from the fact that *Time*'s vocabulary seems to grow more colloquial in general in this period (at least in our eyes, e.g. *guy*, *lot*, *thing*), a number of controversial taboo subjects seem to have become discussable: *sex*, *drug*. 'Terrorism' and 'intelligence' seem to have become major concerns in post-09/11 America, and perhaps the presence of the word 'attack' and 'technology' might be (partially) interpreted in the same light. Again, we see how vocabulary related to media absolutely dominates the final rankings in the corpus: Hollywood seems to have enjoyed an increasing popularity (*film*, *director*, *movie*, ...) but it is information technology that seems to have had the biggest cultural impact: mobile communication devices (*phone*, *cell*) and Internet-related terminology (*network*, *computer*, *internet*, ...) seem to have caused a major turning point in Time's lexis.[6]

[6]Due to lack of space, we only report results for $\lambda = 0.1$ applied to nouns, but highly similar results could be obtained

5 Twentieth Century Turning Points?

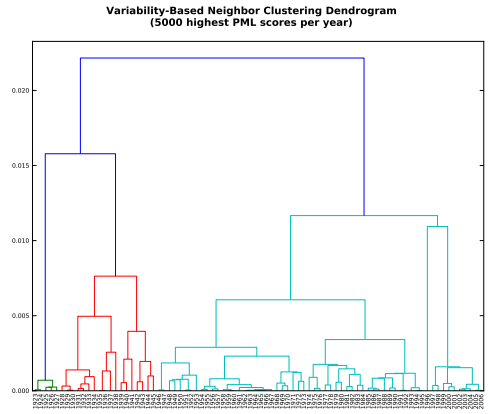
An interesting technique in this respect is a clustering method called VNC or 'Variability-Based Neighbor Clustering' (Gries and Hilpert, 2008). The technique has been introduced in the field of historical linguistics as an aid in the automated identification of temporal stages in diachronic data. The method will apply a fairly straightforward clustering algorithm to a data set (with e.g. Ward linkage applied to a Cosine distance matrix) but, importantly, it will add the connectivity constraint that (clusters of) data points can only merge with each other at the next level in a dendrogram, if they are immediately adjacent. That is to say that e.g. in a series of yearly observations 1943 would be allowed to merge with 1942 and 1944, but not with 1928 (even if 1943 would be much more similar to 1928 than to 1943). We have applied VNC (with Ward linkage applied to a plain Cosine distance table) to a series of vectors which for each year in our data (1923-2006) contained the PML scores of 5,000 words deemed most relevant for that year by the model.

The dendrogram resulting from the VNC procedure is visualised in Figure 13. The early history of *Time Magazine* (1923-1927) does not really seem to fit in with the rest and takes up a fairly deviant position. However, the most attention-grabbing feature of this tree structure is the major divide which the dendrogram suggests (cf. red vs. green-blue cluster) between the years before and after 1945, the end of the Second World War. Another significant rupture seem to be present before and after 1996: the discussion leaves us to wonder whether this turning point might related to the recent introduction of new communication technologies, in particular the rise of the Internet.

Historically speaking, these turning points do not come as a surprise. There is, for instance, widespread acceptance among historians WWII has indeed been the single most influential event in the twentieth century. What does surprise, however, is the relative easy with which a completely unsupervised procedure has managed to suggest

using other part-of-speech categories and settings for λ. An interesting effect was associated with 'fiddling the knob' of this last parameter: for lower values (0.01, 0.001 etc.), the model would come up with perhaps increasingly characteristic, but also increasingly obscure and much less frequent vocabulary. For the fourties, for instance, instead of returning the word 'bomber' the analysis would return the exact name of a particular bomber type which was used at the time. This parameter setting deserves further exploration.

Figure 13: Dendrogram resulting from applying Variability-Based Neighbor Analysis to vectors which contain for each year the 5,000 words deemed most relevant by the PML.



this identification. Arguably, this is where we leave the realm of the obvious when it comes to the computational study of cultural history. The identification of major events and turning points in human history is normally a task which requires a good deal of formal education and some advanced reasoning skills. Here, we might be nearing a modest form of Artificial Intelligence when we apply computational methods to achieve a fairly similar goal. Hopefully, these analyses, as well as the ones reported above, illustrate the huge potential of computational methods in the study of cultural history, even if only as a discovery tool.

6 Conclusion and criticism

In this paper we have discussed a series of analyses that claim to mine a data-driven cultural characterization of the 'Zeitgeist' of some of the main periods in the twentieth century. Nevertheless, we must remain vigilant not to overstate the achievement of these techniques: it remains to be determined to which extent can we truly call these applications Digital History and whether these analyses have taught us anything which we did not know before. Because the twentieth century is so well known to most of us, the evidence often tends to be self-referential and self-explanatory, and merely confirms that which we already knew intuitively. Like with most distant reading approaches, the results urge us to go back to the original material for the close reading of individual sources in their historical context, in order to ver-

ify the macro-hypotheses that might be suggested at a higher level. Therefore, the proposed method might in fact be more suitable for the study of time periods and corpora of which we know less.

Nevertheless, our methodology seems promising for future applications in Digital History: our naïve periodisation in decades, for instance, might be hugely fine-tuned by processing the results of a VNC-dendrogram. Breaking up history into meaningful units is a much more complex, and often controversial matter (e.g. 'When does modernity start?'). In this light, it would be helpful to have at our disposal unbiased, computational tools that might help us to identify cultural ruptures or even turning points in history. Our results reported in the final section do show that this application yields interesting results, and again, the method seems promising for the analysis of lesser known corpora.

Acknowledgments

The authors would like to thank the anonymous reviewers and Kalliopi Zervanou for their valuable feedback on earlier drafts of this paper, as well as Walter Daelemans and Antal van den Bosch for the inspiring discussions on the topic. For this study, Mike Kestemont was funded as a postdoctoral research fellow for the Research Foundation of Flanders (FWO). Folgert Karsdorp was supported as a Ph.D. candidate by the Computational Humanities Programme of the Royal Netherlands Academy of Arts and Sciences, as part of the Tunes & Tales project.

References

Alberto Acerbi, Vasileios Lampos, and Alexander R. Bentley. 2013a. Robustness of emotion extraction from 20th century English books. In *BigData '13*. IEEE, IEEE.

Alberto Acerbi, Vasileios Lampos, Philip Garnett, and Alexander R. Bentley. 2013b. The Expression of Emotions in 20th Century Books. *PLoS ONE*, 8(3):e59030.

Mark Davies. 2013. TIME Magazine Corpus: 100 million words, 1920s-2000s.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stefan Th. Gries and Martin Hilpert. 2008. The identification of stages in diachronic data: variability-based neighbour clustering. *Corpora*, 3(1):59–81.

Djoerd Hiemstra, Stephen E. Robertson, and Hugo Zaragoza. 2004. Parsimonious language models for information retrieval. In Mark Sanderson, Kalervo Jrvelin, James Allan, and Peter Bruza, editors, *SIGIR*, pages 178–185. ACM.

Patrick Juola. 2013. Using the Google N-Gram corpus to measure cultural complexity. *Literary and Linguistic Computing*, 28(4):668–675.

Kalev H. Leetaru. 2011. Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 16(9).

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182.

R Core Team, 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Steve Skiena and Charles Ward. 2013. *Who Belongs in Bonnie's Textbook?* Cambridge University Press.

Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jean M. Twenge, Keith W. Campbell, and Brittany Gentile. 2012. Increases in Individualistic Words and Phrases in American Books, 19602008. *PLoS ONE*, 7(7):e40181.

Gerben Zaagsma. 2013. On Digital History. *BMGN – Low Countries Historical Review*, 128(4):3–29.