Identifying Epochs in Text Archives

Tobias Blanke
Department of Digital Humanities
King's College London, UK
Email: tobias.blanke@kcl.ac.uk

Jon Wilson
Department of History
King's College London, UK
Email: jon.wilson@kcl.ac.uk

***Abstract*—This paper develops an automated approach to the 'distant reading' of textual archives in order to classify epochs in the use of language and examine their particular characteristic. It classifies epochs by applying a series of standardised dictionaries to map the semantics of government documents, using the changing frequency of terms in these dictionaries to identify moments of rupture in language. It then tests a variety of techniques to chart the relationship between the changing shape of individual linguistic elements and aggregate patterns, particularly topic models and word2vec word embeddings. The result are a set of largely automated tools for understanding the structure of digital textual archives.**

***Keywords*—Computational Archives, Digital History, Cultural Analytics**

I. Introduction

Franco Moretti famously called for a quantitative approach to track literature's trends. He invented a 'materialist sociology of literary form' [1]. In our work, we follow Moretti by using the techniques he describes as 'Distant Reading' to develop a materialist approach to political language. The imprecise and variable culture of political language makes it necessary to consider concepts in their aggregates and build models which relate them to their particular contexts. If a collection of texts is time-indexed, a quantitative analysis allows us to trace the changing shape of political language, by tracking clusters of terms relating to particular concepts and - by tracing word embedding - charting the changing meaning of words.

In our previous work, 'The Free Economy and the Welfare State', we used this approach to question dominant narratives about the history of late twentieth-century British politics, posing questions such as when interest in the welfare state peaked. Taking into account a word's contexts, we added to this analysis a change in the meaning of these contexts over time. The word 'safety', for instance, was linked in the 1960s but connected to health issues form the late 1970s. The overall thrust of our argument was to challenge current uses of the concept of 'neoliberalism', offering an alternative account showing that the increasing importance of the free market occurred alongside the growth of an interventionist welfare state [2].

But while successfully distant-reading the political concepts in play during a number of epochs, we did not use distant reading to determine what the epochs themselves were. Our analysis challenged prevalent understandings of how to characterise the politics of particular epochs, but retained a standard periodisation, assuming breaks occurred with changes of government in 1964, 1979 and 1997 in particular.

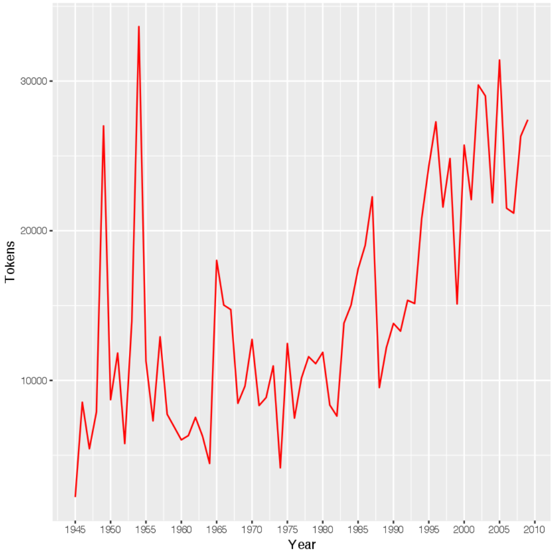

Fig. 1. Distribution of Tokens in UK White Papers

In this paper, we experiment with automated and quantitative approaches to determine the classification of time-coded collections of textual collections into epochs and periods. This method should be useful to many comparable textual collections in archives. We do so by applying a series of standardised dictionaries to map the semantics of government documents according to three generic categories: ambiguity, fairness and morality. This result of this semantic analysis leads us to classify an archive of political texts into three epochs of political communication. We then trace changes of meaning in key political concepts across these three epochs; employing topic models and word2vec word embeddings to do so.

II. Related Work and Data

Moretti's work is generally discussed in relation to the question of traditional 'close' reading of texts or machine but 'distant' reading of texts [3]. Jockers prefers terms micro- and macro-analysis instead of close and distant reading [4] to emphasize the use of statistical methods to analyse text archives. Despite such influential voices, distant reading remains controversial in the (digital) humanities and its related computational archival work [5], which are more comfortable with the close analysis of individuals and their relations or patterns of argumentation. Nevertheless, there are now more and more examples of distant reading of textual archives, demonstrating that distant readings of texts can result in a unique description of the structure of the underlying collections and using all textual features available, including metadata, machine-read relationships or statistical analysis of word correlations. The result is a better descriptions of the structure of textual archives [6].

In this paper, we focused on UK Government White Papers to map connections and similarities in political communications from 1945 to 2010. These are 888 documents, 19.3 million words in total. Figure 1 presents the average number of tokens/words had in a particular year. There is a clear trend towards longer and longer documents as well as several early outliers, which would merit further investigation. White papers are a good starting point for a materialist analysis of political communications because they represent an official, public statement of government policy on particular topics. They stand at a sort of mid-level of abstraction, between theoretical debates by intellectuals and intellectually-minded politicians and the texts which produce action themselves: as laws, regulations, guidance documents, instructions and so on. Consequently, they not only reflect what politicians and officials think about a particular topic, but offer a good sense of how they think of practical political action itself is thought about: of who should do what to what or whom.

III. Detection of Epochs of Political Discourse

In the first stage of our analysis, we cleaned and created a united corpus for all of the collection using the quanteda package in the R language [7]. We create automated semantic annotations for three annotations using dictionaries: ambiguity, fairness and morality. Quanteda makes it easy to import dictionaries from several formats commonly used in the humanities and social sciences. Dictionaries are high-quality linguistic resources that allow for automated annotatations and content analysis, which we use to provide a preliminary classification of textual content according to topics in the texts. This is, however, a highly simplified approach to semantic annotation and not without its critics [8]. Nevertheless dictionary-based semantic annotations are often the only real option in humanities and social science where alternatives such as supervised learning of text classifications are not possible because of the lack of reference training collections.

Dictionary-based approaches are based on a word-frequency analysis where the reference tokens/words are limited to keywords in the dictionary. Words in the dictionary map textual content to specific topics and categories, the degree of which depends on the number of words that belong to each topic. The approach is simple and effective by first identifying all dictionary-based keywords in a document and then filtering all those that belong to categories.

We ran this approach with three dictionaries for four annotations. The first dictionary maps different moral concepts, based on the effort of the social psychologist Jonathan Haidt and colleagues typologies of different kinds of morality [9]. Their dictionaries separate moral concepts into distinct domains associated with evolved adaptations for living in social groups [10]. We kept the categories covering general moral terms and fairness. The general moral category included terms like ’bad’, ’character’, ’correct’, etc. Fairness listed terms such as ’balance*’, ’constant’, ’egalitar*’, etc.

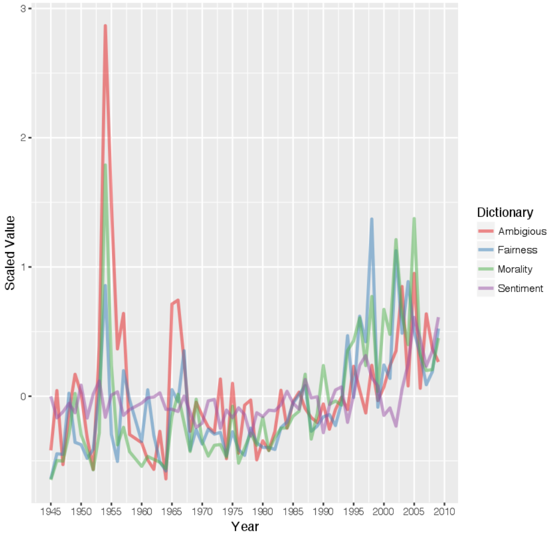

Fig. 2. Dictionary-based Annotation of UK White Papers

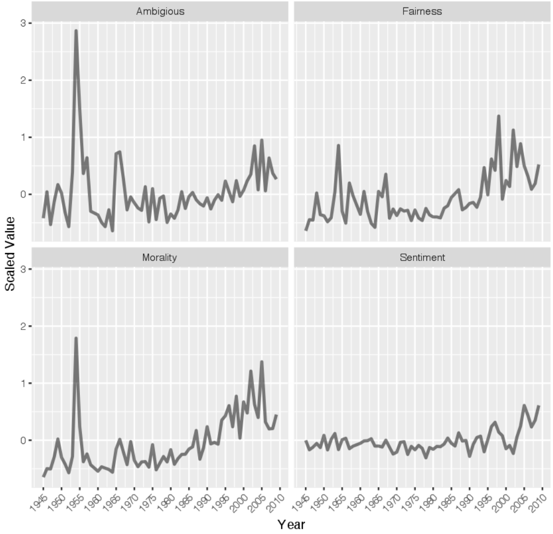

Fig. 3. Facetted Annotations of UK White Papers

The second dictionary uses sentiment vocabularies to map policy developments [11]. The final dictionary measured political ambiguity [12]. This dictionary uses three classifications of writing quality in English texts. We kept only the ’Ambiguous designation’ class containing terms such as ’whatever’ or ’no particular’.

All semantic annotation values, as provided by the dictionaries, were normalised and scaled to center around the mean value. Negative values then indicate a lower than average score and positive value a higher than average one. Finally all value are indexed with their years and the average of all categories per year determined. Figure 2 brings together the resulting trends into one graph, while Figure 3 splits the same graph into three sub-graphs to emphasise the developments per annotation: ambiguity, fairness, morality and political sentiment.

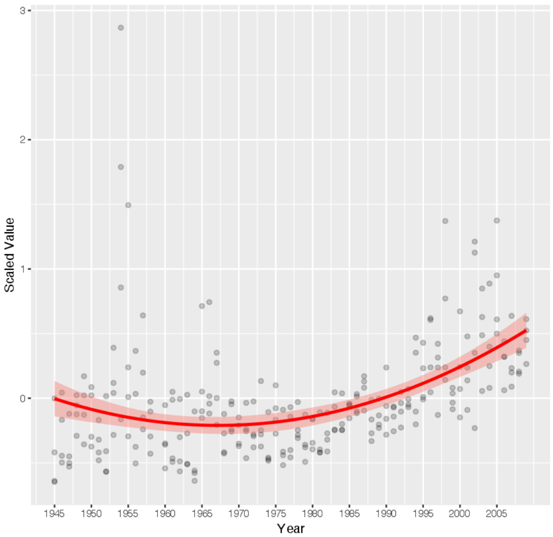

Fig. 4. Quadratic Regression of UK White Papers Annotations

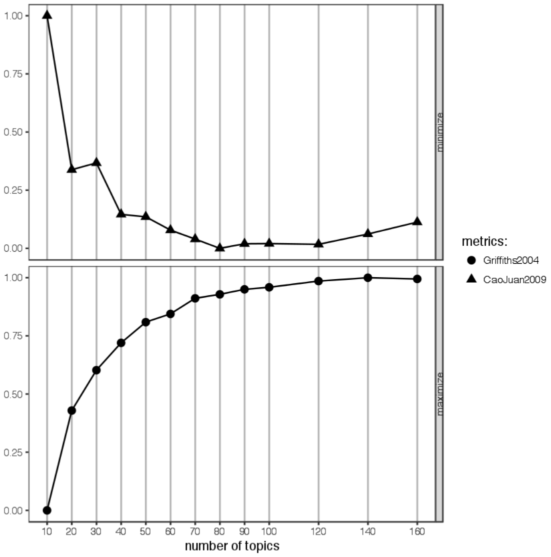

Fig. 5. Optimal number of topics

We share the critique of dictionary-based approaches that they are not exact enough for single observations. More work needs to be done to determine the nature of individual spikes, for example, determining the exact kind of vocabulary they represent or the number of words involved - among other things. Instead of trusting individual results from the dictionary-based analysis, we rather use them to determine trends and points of change in political communications, aggregating the results of a number of dictionaries. To make the case stronger, we use three relevant dictionaries rather than a single one, as it is common in, for instance, sentiment analysis [8].

We should stress that we are not necessarily convinced of the methodology underpinning any of these dictionaries. In our analysis they are simply collections of similar types of words, whose frequency would remain constant if texts had a similar style and subject matter; but which would change when the structure of political change altered. In other words, we use them to indicate moments of rupture, not for their content. Accordingly, our approach could be repeated with other dictionaries.

Examining Figures 2 and 3, it seems that at the beginning there is a period of highly fluctuating political communication before the scaled communication annotation values turn generally negative in the mid 1960s before emphatically turning positive after the 1990s. Figure 4 formally analyses these trends and turning points. The red line represents a quadratic regression modelling all the data points and aggregating them, to allow us to identify moments of change in political language.

Figure 4 summarises our analysis of turning points in political communications. 1965 and 1990 both mark transitions in the history of political language and allow us to define three epochs overall. The first one until 1965 is overall lower than the average for all four topics. It also has a slight negative trend, but not a very strong one. From 1965, this trend turns positive until in 1990 all indicators turn higher than the average. The overall positive trend then also accelerates and becomes more defined.

Thus, we have three epochs in our textual archives. The first runs from 1945 to 1964, while the second starts in 1965 and ends in 1990. The final epoch seems to be from 1990 until the end of our archives in 2009.

IV. Automated analysis of Epochs

The most common approaches to quantitative textual analysis track the frequency of clusters of related terms determined manually, including both the techniques developed by Franco Moretti, and the dictionary approach we used to identify epochs above. The benefits of using our human capacity to cluster terms are clear. The danger, though, is that we miss patterns we would not otherwise notice and instead impose our own current conceptual systems on the archive, in the process undermining the empirical emphasis of materialist textual analysis. In this section, we experiment with two automated techniques for analysing shifts in individual concepts and clusters of concepts, both available within the easily accessible gensim python library. The aim is to begin developing a sense of what exactly changed in the transition between our three periods, and also to verify that the periodisation is correct.

A. *Topic Models*

Our first attempt to understand the three auto-detected epochs involves topics models for each of the epochs, which we then use to create frequency distribution graphs to confirm the importance of particular clusters in each epoch. We use Latent Dirichlet Annotation (LDA), a popular topic analysis approach [13], which is, however, also very resource-intensive. It is commonly used to extract topics from textual data and study patterns. LDA assumes that each document belongs to a number of topics (k) with a certain probability. Assuming that we have n documents, LDA produces a matrix of k x n, describing the probability of each document belonging to certain topics. The number of topics k needs to be chosen

Epoch	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
All	primaiy	cynulliad	renewables	ratepayers	mecu	bnfl	tabic	swansea	cymru	flat-rate
	seivices	bydd	biomass	amalgamation	colonial	decommissioning	docklands	newport	addysg	widow
	industiy	llywodraeth	decc	mccs	kenya	ukaea	efls	glamorgan	gyfer	pensioner
	sendee	cymru	tidal	rateable	nigeria	magnox	million	monmouthshire	nghymru	earnings-related
	seivice	gymru	biofuels	tsbs	obligatory	nirex	attendances	valleys	hyfforddiant	solvency
1945-1964	output	health	education	sterling	scotland	article	defence	soviet	tons	korea
	growth	bill	courses	loans	scottish	federation	army	berlin	food	conference
	investment	works	research	payments	wales	constitution	forces	states	average	nations
	industries	councils	schools	table	road	commission	military	germany	consumption	commission
	labour	medical	broadcasting	gold	building	federal	atlantic	governments	wheat	korean
1965-1989	curriculum	tunnel	kong	presidency	hectares	spanish	subsection	devolved	patients	bypass
	ethnic	heathrow	hong	spain	crop	gibraltar	clause	elections	cable	online
	deregulation	airlines	chart	directives	breeding	dispersal	chargeable	ballot	offenders	junction
	absent	stansted	contingency	portugal	wheat	spain	solicitors	museum	patient	motorway
	rebate	gatwick	revalued	gatt	potatoes	majestys	disciplinary	devolution	custody	competent
1990-	subsection	landfill	charters	smoking	seafarers	bydd	cymru	rdas	census	para
	clause	wastes	subsistence	pharmacy	nato	mewn	gyfer	assemblies	chamber	airport
	divorce	radioactive	governments	genetic	ship	gyfer	gwariant	councillors	pupil	mayor
	confiscation	renewables	museum	genetics	judgment	cymru	swyddfa	hong	learners	crossrail
	registrar	heat	superannuation	pcts	euro	cynulliad	gwasanaethau	psas	anti-social	tecs

TABLE I. TOPICS PER EPOCH

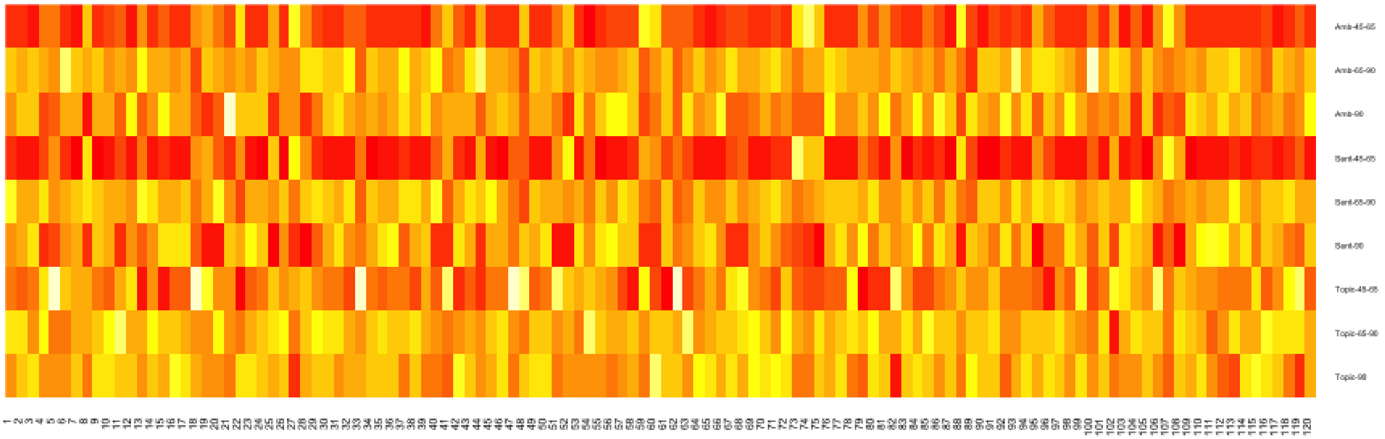

Fig. 6. Visualisation of annotation for topics

in advance. There is no single way of choosing the correct number of topics. A commonly used method is the brute-force approach of fitting the topic model for a range of different topic numbers. Figure 5 shows the result for the whole corpus based on three standard topic modelling evaluation measures assessing maximising likelihood and minimising Kullback-Leibler divergence [14]. It seems that the best number of topics is in the range of 90 to 130. This is confirmed by the metadata of our data: each document has been humanly annotated by topics, 'education', 'health', 'scotland', 'defence' and so on, with 97 different topics in this column.

Another important input to LDA is the choice of concentration parameters: α and δ. High α values mean documents belong to many topics. Higher δ values mean the topics contain many words. We follow convention and set $\alpha = 50/k$, where k is the number of topics and $\delta = 200/m$, where m is the total number of unique features (words) in the documents. We specify that words need to have a minimum frequency of 25 in the collection and appear in at least 50 documents. This leaves us with 7,500 features and $\delta = 0.027$ as well as $\alpha = 0.42$ with $k = 120$. Again, there are other more fine-grained methods to choose the concentration parameters, which we would use in a more detailed analysis of the topics. Here, we concentrate on determining topic changes between the epochs.

We continue with the pairwise Pearson correlations between the LDA topic clusters and the values for ambiguity, fairness, morality and political sentiment for each epoch in relation to the whole corpus. Figure IV-A is the visualisation of the correlations between annotations and 120 topics in the corpus. The heat map's cells are darker and redder the stronger the correlation. The pattern of correlations reveals, for instance, that ambiguity (Amb) between 1945 and 1965 is strongly correlated with many topics and so is Sentiment (Sent) in the earliest epoch. But overall there are too many topics to make this analysis really useful for investigating individual correlations. Table I contains the first 10 topics of 120 for the whole collection under the 'All' tab.

Figure 6 is useful for an ad-hoc evaluation of the epoch determinations. For all three recorded annotations - ambiguity, sentiment and topics - there are strong transitions in the importance of individual topics for all epochs. The colour intensity changes in all columns of the heat map. For instance, for ambiguity the first cell for 1945-9165 is dark red, for

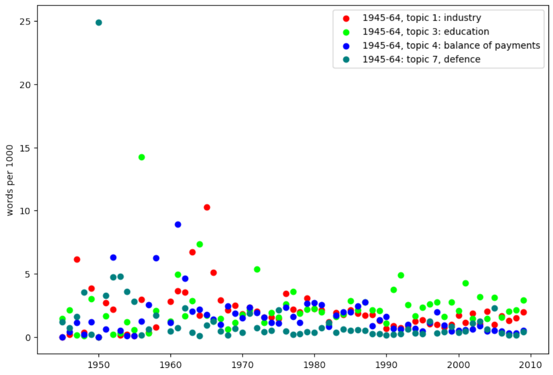

Fig. 7. Frequency of topic clusters, 1945-64

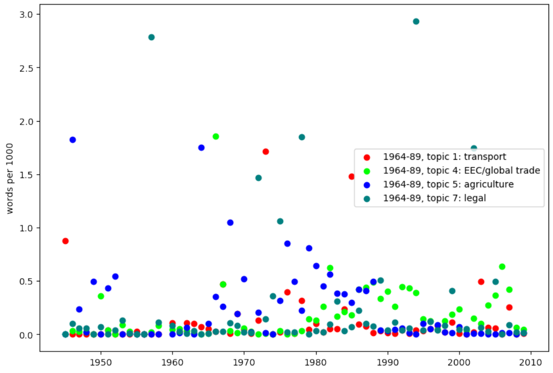

Fig. 8. Frequency of topic clusters, 1964-89

1965-1990 is light orange and then gets post 1990 darker orange. This means, that there are topic shifts correlated to the annotations between the epochs, which makes us confident that they were correctly identified.

One note on final column 'Topic'. It should be more strongly correlated to the automatically generated topics than it is. The reason for this disagreement is that we increased the number of topics to 120 rather than the manually determined optimum of 97.

Table I provides an overview of the top 10 topics for each epoch. It is worth noting that from the 1990s the Welsh language starts appearing in the texts; we see a distinctive set of clusters for each epoch. We can start to use the English clusters to map changing political themes. Industrial production comes first (topic 1). A number of clusters in the first epoch show the important of the U.K.'s post-war geopolitical position, with an emphasis on the cold war conflicts and defence (topics 7, 8 and 10), on the need to maintain a positive balance of payments (topic 4), and on the constitutional process of decolonisation (topic 6). In the second epoch, we see a turn towards transport infrastructure (2 and 10), and agricultural production (topic 5). Geo-political concerns have shrunk in scale, with two clusters concerned with small overseas territories left over from the process of decolonisation, Gibraltar and Hong Kong (topics 3 and 6). Transport infrastructure continues into the third period, 1990-2009 period (topic 10), but we also see discussion of political devolution (topic 8), public health (4), and environmental concerns (topic 2).

Topics are distinctly different between the epochs, and also relatively coherent, confirming our epoch detection. To further confirm the link between particular topics and epochs, we plotted the frequency of terms in each cluster across our archive's full temporal span, in the process creating a more fine-grained account of the relationship between concept-clusters and epochs. Figures 7, 8 and 9 show the frequency of terms for four of the most intuitively clusters computed from each epoch across the whole time period. These graphs confirm our periodisation, by showing that the frequency of terms generated in particular epochs concentrate in that epoch although correspondence is not exact.

B. *Word2Vec for Epochs*

Our approach thus far has focused on tracking the frequency of clusters of individual terms. As with many natural language processing techniques, these approaches treat words as atomic units. To understand the changing shape of meaning in a series of texts we also need to trace changes in the relationship between words. The word2vec model developed by Tomas Mikolov and his colleagues at Google allows us to do exactly that [15]. Word2vec gives each term a position in a multi-dimensional vector space, and thus enables us to build a model representing the linguistic contexts of each word. It then allows us to discern the concepts associated with a particular term by using cosine similarity to calculate the most similar terms.

We began by creating word2vec models for each of our three epochs. To discern where meaning had changed the most, we computed the terms whose semantic context had changed the most across any two epochs. The method worked best with relatively small overall vocabulary lists. The following results used the 500 most frequent words contained in the epochs compared in each case; larger lists included too many words that weren't present in enough contexts for meaningful comparisons. We experimented comparing lists of most similar words of differing lengths, but found there was little difference in results as long as the list was more than 50 and less than 500. We compared word2vec models between 1945-1964 and 1965-1990, and between 1965-1990 and 2009.

The 20 terms which changed meaning the most between 1945/64 and 1965/1990 were as follows: 'korean', 'gold', 'berlin', 'home', 'reserved', 'u.k.', 'federation', 'broadcasting', 'federal', 'currency', 'company', 'legislative', 'conference', 'balance', 'association', 'war', 'great', 'Korea', 'information', 'ltd'. Many of these words come from the discussion of the UK's foreign relations, indicating that some of the greatest changes between these periods occurred both with the shifting international situation and post-imperial Britain's changing place in the world. In the earlier period 'Berlin' occurred in the context of post-war reconstruction and the cold war (associated terms are zone, occupation). After 1965 it simply becomes another city, occupying a similar place as Washington or Stockholm. Between 1945-64, 'war' was an active process the British state was involved with or immediately recovering from, located close to words concerned with things happening in war ('prisoners', 'damage' and 'internees') or the dynamic occurrence of conflict ('short', 'long', 'outbreak'). In the second period 'war' was memorialised and compensated far more than fought, with 'graves', 'memorials', 'death', 'pensions' 'widows' and 'retirement' commonly embedded terms. Contexts for other terms show the shift in attitudes amongst policy-makers from war and international relations to domestic concerns. For example, 'settlement' started off referring to the resolution of international conflict ('cessation', 'territory', 'peaceful'), but was associated the resolution of domestic disputes in the second epoch.

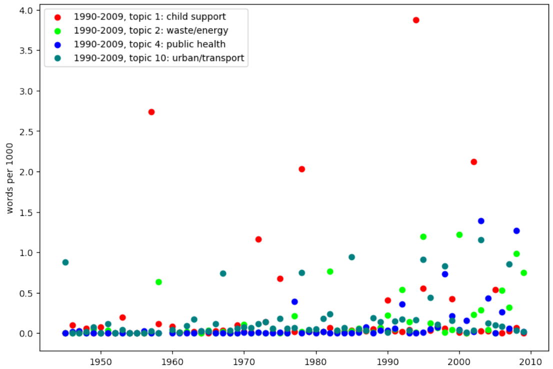

Fig. 9. Frequency of topic clusters, 1990-2009

Turning to the second transition, between 1965-1989 and 1990-2009, the largest group of words whose embeddings change the most concern the governance of the national economy. The 20 terms with the greatest change in embeddings between the two epochs are: labour', 'manpower', 'nationalised', 'building', 'ireland', 'current', 'britain', 'northern', 'outturn', 'kingdom', 'house', 'industrial', 'improvement', 'administration', 'survey', 'community', 'states', 'third', 'about', 'home'. We see here the decline in explicit references to the process of production. For example, 'labour' is associated with a series of synonyms between 1965-1989 ('craftsman', 'manpower', 'worker', 'staff', 'manpower'), but is linked with a less coherent bundle of terms in the last period, indicating the decline of the word in political discourse. 'Building' refers to physical construction between 1965-1989 ('works', 'renovation', 'housebuilding', 'construction') but develops a more metaphorical use, becoming associated with broader processes of creation ('creating', 'acquisition', 'focusing').

With this focus on the transformation in word relationship, using word2vec offers the possibility to explore the characteristics of each epoch. By comparing embeddings for the same word across different epochs, and then ranking those words, we have created an unsupervised, automated method for discerning which changes are most significant between each period. There is scope to develop such an approach both with this corpus, and other bodies of texts.

V. Conclusion

This paper has developed a 'materialist sociology of political texts' of post 1945 UK government white papers. Compared to our earlier attempts, we relied on machine-reading techniques for texts not only to read the texts themselves but to develop ways of classifying of epochs. We applied a series of standardised dictionaries to annotate the government documents according to their ambiguity, fairness, morality and political sentiment. This result of this semantic analysis leads us to classify an archive of political texts into three epochs of political communication. We then traced changes of meaning in key political concepts across these three epochs, testing two techniques, topic models and word2vec word embeddings. Each approach validated the strong differences in political communication in the three epochs, allowing us to begin developing a sense of strong themes in each.

This approach is useful not just to detect changes in political communications but also shifts in other textual archives. Given the right combination of dictionaries we can detect epochs of interest, and then use topic modelling and - in particular - word2vec word embeddings to analyse the details of the conceptual shifts using automatic techniques.

The techniques developed here add to the tools which particularly contemporary historians have at their disposal in analysing recent historical eras which have left voluminous, relatively easily accessible archives. 'Distant reading' offers the possibility of tracing aggregate patterns in the archives, instead of using intuition or recent historical prejudice to decide which fraction of the archive to read. Our method of classifying epochs allows us to examine change through time, with the ultimate aim of creating a richer, more empirically valid understanding of the recent past.

References

[1] F. Moretti, *Graphs, maps, trees: abstract models for a literary history*. Verso, 2005.

[2] J. Wilson, "The free economy and the welfare state," 2017.

[3] S. Ross, "In praise of overstating the case: a review of franco moretti, distant reading," *Digital Humanities Quarterly*, vol. 8, no. 1, 2014.

[4] M. L. Jockers, *Macroanalysis: Digital methods and literary history*. University of Illinois Press, 2013.

[5] J. Goodwin, J. Holbo, and F. Moretti, "Reading graphs, maps, and trees: Responses to franco moretti," 2011.

[6] S. Jänicke, G. Franzini, M. F. Cheema, and G. Scheuermann, "On close and distant reading in digital humanities: A survey and future challenges," *Proc. EuroVis, Cagliari, Italy*, 2015.

[7] K. Benoit and P. Nulty, "quanteda: Quantitative analysis of textual data," *R package*, 2016. [Online]. Available: https://github.com/kbenoit/quanteda

[8] J. Grimmer and B. M. Stewart, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political analysis*, vol. 21, no. 3, pp. 267–297, 2013.

[9] J. Haidt, *The righteous mind: Why good people are divided by politics and religion*. Vintage, 2012.

[10] J. Haidt and J. Graham, "When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize," *Social Justice Research*, vol. 20, no. 1, pp. 98–116, 2007.

[11] L. Young and S. Soroka, "Lexicoder sentiment dictionary," *McGill University, Montreal, Canada*, 2012. [Online]. Available: lexicoder.com

[12] J. H. Hiller, D. R. Marcotte, and T. Martin, "Opinionation, vagueness, and specificity-distinctions: Essay traits measured by computer," *American Educational Research Journal*, vol. 6, no. 2, pp. 271–286, 1969.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[14] N. Murzintcev, "ldatuning: Tuning of the latent dirichlet allocation (lda) models parameters," *R package*, 2015. [Online]. Available: https://github.com/nikita-moor/ldatuning

[15] T. Mikolov, K. Chen, G. Corrodo, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv*, 2013.