Systematically Detecting Patterns of Social, Historical and Linguistic Change: The Framing of Poverty in Times of Poverty

Schneider, Gerold

DOI: https://doi.org/10.1111/1467-968X.12252

Systematically detecting patterns of social, historical and linguistic change: the framing of poverty in times of poverty

ABSTRACT

The linguistic DNA project seeks to understand the evolution of philosophy, society and language during the Modern English period. Corpora like Early English Books Online (EEBO), Corpus of Late Modern English Texts (CLMET) and Corpus of Historical American English (COHA) allow us to apply statistical data-driven models extracting patterns to confirm our expectations. As systems biology has revolutionised biology by systematically searching for all patterns, we detect patterns in our data systematically with contextual and distributional semantic approaches, an approach that could be called systems history.

We uncover semantic patterns with methods from text mining, computational linguistics, and digital humanities. We normalise the spelling automatically to present-day variants and use bottom-up analyses to step from words to concepts: collocations, topic modelling, and distributional semantics.

We illustrate the approaches with two case studies: associations of poverty changing across time, and Charles Dickens social criticism, his vision of helping to improve the situation of the poor.

As no gold standard for our task exists, our approaches are exploratory, which entails considerable manual intervention, e.g. sifting candidate lists, reading excerpts and interpreting topic models. A fully automatic approach is currently neither feasible nor envisaged: semi-automatic approaches give researchers the inspiring opportunity to interact with the texts in a constant move between distant and close reading. The different characteristics of the various statistical methods offer complementary perspectives.

KEYWORDS: data-driven, statistical models, text mining, machine learning, collocations, topic modelling, poverty, Digital Humanities

1 Introduction

This study sets out to show how data-driven computational approaches to historical linguistic corpora can systematically detect changes in society, history, thought styles and language. We aim to detect changes at the level of words, topics and semantics, with the help of two related case studies.

Diachronic linguistics and history have been the subject of investigations of innumerable studies. Why would a data-driven approach be needed? First, there is always a risk of oversight: could one rely on a method which verifies our findings, based on the vast textual resources that are available for some languages, including English? Second, the question how writers of the time framed historical events and which underlying assumptions they took are not necessarily answered. When one reads historical sources, one does so based on our assumptions and current scientific knowledge. As noone has the time to read all the relevant sources, insights are sometimes based on single sources rather than systematic quantitative comparisons. This brings us to our third point: it is hard to get an overview of cultural trends, and put things into perspective. Can we re-study and deconstruct history in a systematic way, which places the voices of the past that we have, the data, in the centre, without pre-conceptions? Or will the reliance on quantitative tools just introduce a new bias?

In many domains of pure science such systematic assessments are under way. The hypothesis-driven research method is increasingly supplemented by data-driven approaches. Ananiadou et al. (2006) state the following:

> Systems biology is one of the key examples of a field where the mode of scientific knowledge discovery is shifting from a hypothesis-driven mindset to an integrated holistic mode that combines hypotheses with data. (Ananiadou et al. 2006: 571)

The argument for a holistic view in history and linguistics is crucial – in some ways even more crucial than in the hard sciences, because we cannot observe linguistic structures under the microscope. The data-driven method of a systematic search for patterns, so-called 'mining' is a popular method in data science. One can use raw data from experiments collected in large databases, or textual data:

> In the data-rich but hypothesis-poor sciences, including functional genomics and most of biomedicine, the normative hypothesis-driven, deductive scientific method becomes increasingly difficult to sustain … As a complement to hypothesis-driven deductive science, we are now witnessing the emergence of data-driven inductive methods of scientific discovery. These are characterized by the rapid 'mining' of candidate hypotheses from the literature (Ananiadou et al. 2006: 571)

Data-driven methods are also known in media content analysis (Schwartz and Ungar 2015) and corpus linguistics, where they are called corpus-driven (Tognini-Bonelli 2001). Tognini-Bonelli points out that the advantages of choosing a data-driven approach are that it needs fewer theoretical assumptions; this is also an advantage in terms of the principle of elegance (Aarts 2019) as t can bring previously undetected

relations and areas of gradience to the surface. A disadvantage is that it is very sensitive to data: "since the information provided by the corpus is placed centrally and accounted for exhaustively, then there is a risk of error if the corpus turns out to be unrepresentative" (Tognini-Bonelli, 2001: 88). In the words of Ananiadou et al. (2006: 572), "[h]ypothesis generation in TM [Text Mining] relies on the fact that 'chance' connections or associations between disconnected entities or facts can emerge to be meaningful."

Finding such connections, which are candidates for expressing general trends quantitatively for the history of thought, has been coined the research area of culturomics by Michel et al. (2010). They motivate the desire for a large-scale quantitative approach as follows:

> "Reading small collections of carefully chosen works enables scholars to make powerful inferences about trends in human thought. However, this approach rarely enables precise measurement of the underlying phenomena." Michel et al. (2010: 176).

As data, they use a large sample of books collected by Google, which comprises 4% of all books ever printed. This sample is also known as the Google N-gram viewer which can be queried online.[1] The viewer provides a tool for conducting culturomics pilot studies to any interested user. Increases, decreases, spikes and comparisons of variant forms can give us preliminary insights both for linguistic and cultural studies.

While Google N-gram viewer is a useful and easily accessible research tool, we reach beyond it in the present article. Our approach is less large-scale, but more data-driven. Although the potential of the method has been recognized in historical corpus linguistics,it is not often used.

An attractive potential of quantitative corpus-based methods that has yet to be fully realized in diachronic studies lies in exploratory, bottom-up approaches (Gries 2011). The label 'bottom-up' stands for a set of techniques in which the data are processed statistically in order to discover structures that had not necessarily been anticipated by the analyst (Hilpert and Gries 2016: 44).

Another important distinction of our outlined approach compared with observing peaks in Google N-grams is that our statistical models have numerous advantages over descriptions, including the fact that they can be multifactorial. This allows us to assess the importance of a wide range of features, make predictions, smooth over random events, and detect outliers.

In a nutshell, we make the case for a new research paradigm, which in analogy to *systems biology*, could be called *systems history* and *systems diachronic linguistics*. This coinage would also fit in with the Linguistic DNA project, which borrows its metaphor from genetics, an area of biology where systems biology plays a key role. The overarching research question (RQ1) is whether the data-driven methods manage to detect meaningful, interpretable and consistent patterns. We apply these methods to two interrelated pilot studies to address further research questions: can patterns,

[1] https://books.google.com/ngrams

topics and associations of poverty be detected in historical language corpora (research question 2)? We have chosen a case study on poverty because this has been a constant challenge for the majority of the population until recently. While we as Modern humanists frame poverty through the lens of Charles Dickens' social criticism, how did the contemporary writers, from 1470 to 1910, frame poverty (research question 3) ? And were Dickens' views really different, visionary, and are they reflected in the data and detected by corpus-driven approaches (research question 4) ?

2 Data

We use the following corpora. They cover a long time period, and were compiled for various purposes.

2.1 The EEBO Corpus

Early English Books Online (EEBO) is not a balanced corpus but a collection of available books printed between 1470 and 1710[2]. It covers about 135000 books. Over 30000 of them were XML-TEI formatted in a collection called EEBO-TCP Phase 1. The material can be downloaded for free. As EEBO is not balanced, but contains far more material in the later than in the early periods, we have used a more balanced random selection, containing 40 million words (Schneider 2020). In the earliest decades, it includes all texts, and then increasingly fewer, down to 20% in the latest periods. We mapped spelling automatically to PDE variants using VARD2 (Baron and Rayson 2008). For some of our experiments, we have split this EEBO sampler into two periods – Early (1470-1599) and Late (1600-1699).

2.2 The CLMET corpus

The CLMET corpus (Corpus of Late Modern English Texts) was compiled by Hendrik de Smet and Jukka Tyrkkö. It is distributed for free from the University of Leuven[3]. CLMET contains over 34 million words, drawn from 6 coarse genres. It is composed of 3 periods of 70 years each, and covers material from 1710 to 1920. It is a clean and carefully POS-tagged corpus.

2.3 ARCHER

ARCHER (A Representative Corpus of Historical English Registers; Biber et al., 1994) is a major resource for studying register variation and diachronic change in British and American English, from the 17th to the 20th centuries. It is balanced in terms of genres, regional variety (British and US) and covers four centuries. Its earlier centuries are available with spelling mapped to PDE (Schneider et al. 2017).

[2] https://eebo.chadwyck.com/about/about.htm
[3] https://perswww.kuleuven.be/~u0044428/

3 Methods

Our data-driven methods detect patterns automatically as a diagnostic step. The interpretation of automatically detected patterns remains an important and partly manual step: the researcher needs to sift lists of candidates and interact with the texts as not all the reported intricate patterns are meaningful. In the domain of Information Retrieval, the distinction between corpus-based and corpus-driven techniques has been described as follows: “Hand-driven techniques tend to be more accessible, theory-driven, abstract, and able to handle small datasets, while data-driven tend to be more transparent, capture more connections, and are able to yield unexpected associations.” (Schwartz and Ungar 2015: 80). Their argument is that hand-driven techniques are based on an abstraction or theory by researchers, and thus run the risk of being further removed from the actual data.

As Gries (2010) points out, corpus-driven approaches hardly ever start with a *tabula rasa.* Well-established tools such as word- and sentence-level tokenizers, part-of-speech taggers, often also syntactic frames and alternations (Schneider 2014), are typically relied on. They are often also essential: without mapping historical spelling variants to their PDE equivalent (so-called spelling normalization), many of our analyses would be useless. For spelling normalisation we use VARD2 (Baron and Rayson 2008). For an evaluation on texts from our target period see Schneider (2020).

The distinction between corpus-based and corpus-driven is thus gradual, with corpus-driven offering a much vaster space of results to be reaped from a corpus than a simple *yes* or *no* to a predefined hypothesis.

While words do not often directly express the meaning, concepts and topics contained in the text, the semantics of what is said can be partly extracted, for example, using distributional semantics (Baroni and Lenci 2010). The framing of topics, i.e. in which light facts are presented, is also correlated to textual patterns (Entman 1993).

3.1 Statistical Models

While the use of statistical models is commonplace in empirical science, they are also increasingly used in linguistics. For example, Evert (2006) and Gries (2010) advocate the use of multivariate regression models. Multivariate models, compared to monofactorial tests such as significance tests, have the advantage that they take multiple factors into account. They are required to weight and measure the various factors that may be involved. Unless a researcher conducting a significance test has independent evidence that the factor he or she is testing is by far the most important one, significance tests have very limited validity. The likelihood that the single tested factor may just be a side-effect of another more important factor is otherwise very high, and it is almost certain that other essential factors are missed.

Advantages of multifactorial statistical models are, among others:

- They need few theoretical assumptions.
- They weight and assess the significance of and interaction between the factors, unlike in a monofactorial perspective.
- They allow one to make predictions. This allows us to predict the future as in meteorology, to attain classifications mirroring speaker choices in an envelope

of variation, to disambiguate, to assess what speakers or listeners find difficult to process, etc.
- Predictions become better than linear combinations due to appropriate factor weights. These factor weights can also be used to rank the factors.
- Interactions between factors can be assessed.
- They can discriminate between random fluctuations and significant differences, e.g. by using cross-validation or repeating experiments with near-similar conditions to test their robustness.
- They reduce the data to simpler, smoother, more abstract and interpretable approximations. While the so-called *data loss* looks like a flaw at the first sight, it is in fact a virtue, a practical implementation of Occam's razor.

Despite their advantages, multifactorial statistical models also have many restrictions. They are typically an intentional reduction of reality into an operationalized system. The main problem is that the correlations they report cannot be interpreted as cause-and-effect *per se*.

> Observed data alone never give us enough information to allow us to draw causal inferences. Results from descriptive models are, in fact, observed information, albeit disguised in mathematical form. They cannot, by definition, be wrong in themselves (except in situations of bad data or bad algorithms) and do not require any assumptions for their validity. Therefore, descriptive models are desirable because, having no unverifiable assumptions, they are indeed the only thing we can study as social scientists. At a minimum, they provide the foundation from which we can advance to higher levels of knowledge (Xie 2011: 345).

Causal explanations need careful interpretations by the scientist and are often not verifiable. A second problem is that the suggested correlations are not necessarily easy to interpret, and can indeed stem from coincidences in the data. We do not advocate naïve interpretations of data-driven science (Anderson 2008).

> However, the call for a descriptive interpretation … should not be taken too far. While any well-defined summary statistic … is always "correct" descriptively, its interpretation does not need to be, and often is not … An appropriate interpretation of "descriptively accurate" … summary statistics is far more difficult than most researchers realize. A great deal depends on the concrete research setting, particularly the research question. (Xie 2011: 345).

While data-driven methods may start with an exploratory, hypothesis-free stage, in the ensuing empirical cycle between data and theory, the hypothesis-forming and hypothesis-testing stages mutually inform each other, in a potentially repeated dialectic process. A frequently used way to test the hypotheses suggested by quantitative methods summarizing large amount of texts, so-called *distant reading* (Moretti 2013) is then to read a subset of the relevant texts in traditional *close reading* to validate, refute or revise the hypotheses. Bearing these caveats, advantages and restrictions in mind, we can now turn to the individual techniques.

3.2 Collocations

Collocations are the prototypical research area for data-driven approaches, as Glynn (2010) summarises. At the same time, he also points out that there are many others:

> Many linguists believe that corpus-driven research is restricted to the study of collocations … Although the study of 'words and the company they keep' lies at the origins of corpus linguistics, methods for studying corpus data have developed much since those times, both in terms of the kinds of questions that are asked and the kinds of techniques used to answer them (Glynn 2010: 10).

There is a wide selection of association measures: see Evert (2009) for a summary and Pecina (2009) for an overview. We use a variant of the easily interpretable measure O/E, where O is the co-occurrence of two words from a corpus, and E is the count that one would expect if words (say *x* and *y*) were independent events, i.e. if word order were random. O/E can be calculated as follows. The independent probability of generating *x* is its frequency in the corpus divided by the total number of word tokens in the corpus; and for *y* analogously. The probability of *x* and *y* in combination, in other words the observed value (O), is the frequency of *x* and *y* in combination (e.g. the first word in the bigram is *x*, the second *y*) divided by the total number of word tokens in the corpus.

$$p(x) = \frac{f(x)}{N};\;\; p(y) = \frac{f(y)}{N};\;\; p(x,y) = O = \frac{f(x,y)}{N}$$

If co-occurrence of *x* and *y* is due to chance, i.e. if there is no collocational force, then the independent probability of seeing both Expected (E) and Observed (O), the joint probability of seeing the combination, are roughly equal:

$$O = p(x,y) \cong p(x) \cdot p(y) = E$$

O/E, Observed divided by Expected, is then:

$$\frac{O}{E} = \frac{p(x,y)}{p(x) \cdot p(y)} = \frac{\frac{f(x,y)}{N}}{\frac{f(x)}{N} \cdot \frac{f(y)}{N}} = \frac{f(x,y) \cdot N \cdot N}{f(x) \cdot f(y) \cdot N} = \frac{f(x,y) \cdot N}{f(x) \cdot f(y)}$$

As O/E has a tendency to overreport rare events, we use O^2/ E, a variant that is more mixed and has higher accuracy (Bartsch and Evert 2014). As the ARCHER corpus that we are going to use is quite small, we apply a formulation of O^2 that allows us to include a larger corpus, thus achieving smoother results, and incorporate the fact that our expectations are partly based PDE. $O^2 = O*O = O_1 * O_2$ allows the possibility to take two different O's, i.e. observed counts. We take one from the ARCHER period, and one from a PDE corpus. We also use the same smoothing on the expected counts (E). As PDE corpus we use the BNC, so x_{BNC} and y_{BNC} are the counts for *x* and *y* from the BNC.

$$\frac{O^2}{E} = \frac{O \cdot O}{E} = \frac{f(x,y) \cdot f(x_{BNC}, y_{BNC}) \cdot N}{(f(x) + f(x_{BNC})) \cdot (f(y) + f(y_{BNC}))}$$

3.3 Distributional Semantics

Interpreting thousands of collocations can be demanding and confusing, and for a semantic investigation, the fact that synonyms and related words are completely ignored is a serious disadvantage. Collocation detection methods do not automatically aggregate synonyms or semantically related words. While most quantitative linguists are used to exploiting the Firthian hypothesis in order to detect collocations (e.g. Evert 2008, Bartsch and Evert 2014), the collocational method also lends itself to detecting semantically related words or to disambiguating word senses (Schütze 1998). The Firthian hypothesis which literally says that "You shall know a word by its company" (Firth 1957) also holds the key to detecting semantically similar words – words that have very similar contexts tend to be semantically very similar. If one has large corpora of at least several million words, the Firthian method works surprisingly well.

Sahlgren (2006) shows that the size of the context in Firthian approaches plays a crucial role. He notes that using adjacent words (or very small observation windows) delivers classical collocations, i.e. syntagmatic relations, while expanding the window to include increasingly more context delivers results on the paradigmatic axis, synonyms, antonyms, hyponyms (on hyponyms, see Sylvester et al. this volume), associations and related words. This is exploited by the hugely successful paradigm of distributional semantics (Baroni and Lenci 2010, Turney and Pantel 2010). Given large corpora, distributional semantics can achieve surprisingly accurate results. While native speaker accuracy may not yet be within reach, Karlgren and Sahlgren (2001) show that distributional semantics can easily pass the word similarity test of the TOEFL test, thus reaching the level of the intuition of an advanced language learner.

Distributional semantics has come under criticism because it fails to make distinctions between synonyms, antonyms, hypernyms, etc. While this observation is valid from a *prescriptive* perspective, from a *descriptive* perspective, these relations are not axiomatic, and the broad notion of semantic similarity seems perfectly plausible. Miller and Charles (1991) point out that people instinctively make judgments about semantic similarity (including antonyms for example) without the need for further explanations of the concept.

Also the associations that are needed for the analysis of media content and discourse form part of this perspective:

> The kind of co-occurrence marking the relationships among the words that encode the discursive concept belongs to computational distributional semantics. Co-occurrence in this model captures association: a notion of relatedness that is much looser than that captured in formal synonymy (cf. Heylen et al. 2008) or strict collocation (cf. Manning & Schütze 2001, Chapter 5). (Fitzmaurice et al. 2017: 25)

3.4 Topic Modelling

The insights into the relationship between words, documents and contexts have also given birth to linguistic concept modelling (Mehl, this volume) and the method of

topic modelling (Blei 2012). Linguistic concept modelling extends collocations from pairs of two words to triples and quadruples. Collocations include many non-compositional terms and thus help to bridge the gap from word to concept. "Multiword expressions … often offer a better unit of analysis than tokens" (Schwartz & Ungar 2015: 84). Linguistic concept modelling uses large observation windows (Sahlgren 2006) but word triples or quadruples instead of pairs, as in distributional semantics. While it suffers from sparse data even more than collocations do, linguistic concept modelling detects narrow topics which it calls *concepts*. For example, the frequent triple *reason-nature-law* in (Mehl: PAGE) expresses the belief that through reasoning alone we can detect the true underlying laws of nature, and that reason is given by nature and its clear laws can be detected. These almost amount to a definition of scholastic thinking (Taavitsainen and Schneider 2019).

Topic Modelling aims to detect broader topics. It combines document classification with the strong semantic unity of the discourse of a topic and of a document. Topic models increase the context even beyond the level of the document to the topic level. Church (2000) observed that the chance of a word to occur in a document radically changes if the same word has occurred before. While a given word *w* generally has probability p(*w*), the chance of seeing *w* a second time if it has occurred once is much closer to p(*w*)/2 than p(*w*)^2 which we would expect if the first and second appearances were independent. This shift in probability also applies to some degree at the discourse level of topics. Intuitively we know that specific topics are much more likely to generate specific words – this is also the reason why the classification of documents into known classes of topics performs reliably.

Topic modelling optimizes

p(*topic*|*document*) * p(*word*|*topic*)

for all given documents in a collection. It thus combines document classification (p(*topic*|*document*)) and keyword generation (p(*word*|*topic*)). *Documents* and *words* are given, and *topics* are fitted iteratively starting from a random configuration. For our experiments, we are using Mallet.[4]
It can be argued that for well-studied domains, the mapping from words to concepts is available in the form of thesauri, which have been developed carefully and often inspired by painstaking corpus research. Tools mapping words to concepts with rule-based approaches include *WMatrix*[5] (Rayson 2008) for present-day English and the SAMUELS semantic tagger (Piao et al. 2017) for Modern English. While the approach they use performs well on general newspaper texts, specific genres and historical texts often end up with unsatisfactory analyses (e.g. Moreton and Culy 2020), as many of the words have undergone strong semantic shifts or have domain-specific readings or the meanings tend to be expressed indirectly. Addressing the problem of mapping words to concepts by using contexts in a purely data-driven fashion is a viable alternative, provided that sufficiently large amounts of domain data are available. Based on previous experience, about a million words is minimally needed.

[4] http://mallet.cs.umass.edu/topics.php
[5] http://ucrel.lancs.ac.uk/wmatrix/

3.5 Validation

A major challenge in the assessment of the quality of the results that we obtain with topic models is that they cannot be fully evaluated, as no manually annotated correctly labelled dataset (a so-called gold standard) exists. Evaluation partly depends on one's interpretation. Quinn et al. (2010) and Grimmer and Stewart (2013) suggest the following systematic validation for topic models, which we follow in our study: manual verification steps assessing the *semantic* and *predictive* validity should be used. Semantic validity is fulfilled if the suggested topics are internally coherent and distinctive towards the other suggested topics. Predictive validity can be assessed by checking if important events are reflected in the data, typically leading to spikes in the data and to different values in different periods.

4 Results

4.1 Collocations

We first apply collocation statistics to the ARCHER corpus in this section. For this experiment, we have split the ARCHER corpus into an early section, containing the texts from 1600 to 1800, and a late period, from 1900 to 2000. We discarded the texts from 1800 to 1900.

Our collocation method was introduced in section 3.2. We have used a threshold of $f >= 5$. The strongest collocation pairs in each data set are given in Table 1 for the early period, and Table 2 for the late period. Both tables are sorted by decreasing $O*O / E$ in the last column, as defined in section 3.2. O/E and the conditional probability $p(W2|W1)$ are also given. We can see important multi-word entities from each period. We also see how much more multi-faceted today's texts are, technical terms indicating the specialization and fragmentation of our society. We also perceive a shift from religion and the court in the early texts to science and global orientation in the 20th century. The extracts in Tables 1 and 2 only show the top entries of several thousand collocations. Table 1 shows linguistic changes (rank 2-5, 8, 12, 16, 38, 46), the strong influence of religion (ranks 11, 19, 22, 25, 28, 30, 45), famous real or fictional persons (12, 17, 26, 29, 31, 36, 37) and locations (15, 18, 22, 41). In contrast, Table 2 is dominated by specialized technical terms (1, 4, 6, 7, 8, 12, 14, 15, 29, 31, 33, 34, 36, 38, 39, 43) showing scientific and technological progress, further statistics and empirical methods (27) but also specialization and fragmentation of writing and society. The place names are different and more globalized, religious terms have almost disappeared (46). Only four collocations (*Pall Mall*, *New York*, et al. and *Holy Spirit*) appear at the top of both lists. Also political systems (*Royal Highness* versus *Prime Minister*) have changed. Social conventions (*eldest son*, *humble servant*) appear in the early list, but not in the later one, showing that their influence has declined. The world of the 20th century looks very different from that of the 17th and 18th centuries.

Collocations can be evaluated in several ways. One option is to assess if they are non-compositional. Rows 2-5, 8, 13, 14, 16, 27, 38, 44 and 46 in Table 1 are false positives in that perspective. A second option, and arguably more important for our

interest, is to assess if they vary between the two periods, reflecting the social changes that have taken place. This validation is a variant of the predictive validity used by Grimmer and Stewart (2013). From this perspective, the four shared collocations can be seen as false positives.

Looking at the top of the list only shows us the famous tip of the iceberg. Summing from tokens to types was helpful, but we would like to be able to sum up similar types.

RANK	O=F	W1	W2	O/E	CondP	O_{Archer}*O_{BNC} / E
1	11	Pall	Mall	550344.45	0.92	6053788.91
2	40	thou	hast	90320.24	0.04	3612809.68
3	20	thou	shalt	102334.43	0.09	2046688.56
4	7	Thou	shalt	196728.01	0.18	1377096.09
5	15	Thou	hast	76031.36	0.04	1140470.43
6	6	del	Fuego	184789.25	0.05	1108735.47
7	5	inter	alia	221313.92	0.62	1106569.62
8	11	dost	thou	80633.72	0.37	886970.96
9	5	Philosophical	Transactions	166654.6	0.23	833273.01
10	31	La	Motte	23993.95	0	743812.38
11	5	Notre	Dame	148246.03	0.96	741230.14
12	8	Van	Helmont	81649.02	0.01	653192.16
13	12	thou	wilt	44619.45	0.03	535433.42
14	12	hast	thou	43196.64	0.2	518359.65
15	6	Covent	Garden	63307.56	0.99	379845.35
16	8	thou	dost	43982.03	0.01	351856.25
17	44	Royal	Highness	7687.02	0.02	338228.69
18	67	United	States	4779.3	0.39	320212.9
19	17	Holy	Ghost	14469.65	0.03	245984.08
20	37	humble	Servant	6061.59	0	224278.76
21	8	Elector	Palatine	26439.83	0.04	211518.61
22	5	Divine	Providence	41747.87	0.04	208739.34
23	5	Common	Pleas	41084.65	0.02	205423.27
24	12	et	al	16397.78	0.33	196773.37
25	55	Jesus	Christ	3178.14	0.13	174797.61
26	9	Lady	Teazle	18912.24	0	170210.13
27	1436	I	am	114.96	0.02	165083.47
28	7	Holy	Spirit	21100.72	0.34	147705.02
29	6	Daniel	Defoe	24264.45	0.02	145586.72
30	6	NEW	TESTAMENT	23138.92	0	138833.54
31	22	West	Indies	6251.56	0.05	137534.25
32	5	Grand	Vizier	26982.47	0	134912.36
33	6	Lincolns	Inn	21435.04	0.18	128610.26
34	41	New	York	2956.8	0.25	121228.8
35	42	Chief	Justice	2647.05	0.07	111176.19
36	5	Martha	Blount	19767.2	0	98836.02
37	11	Isaac	Newton	8739.86	0.11	96138.48
38	9	wilt	thou	9561.31	0.04	86051.8
39	24	eldest	son	3328.07	0.35	79873.66
40	8	FOR	SALE	9682.1	0.01	77456.78
41	25	de	la	2727.45	0.06	68186.14
42	5	Yearly	Meeting	12864.95	0.13	64324.73
43	144	Sir	John	418.38	0.11	60247.29
44	3584	.	The	16.81	0.09	60246.3
45	8	Old	Testament	7195.15	0.07	57561.17
46	5	shalt	thou	10886.64	0.05	54433.21

Table 1. Top-ranked collocations in ARCHER Early (1600-1800)

RANK	O=F	W1	W2	OE	CondP	O_{Archer}*O_{BNC} / E
1	10	non-steroidal	anti-inflammatory	1211022.05	0.8	12110220.52
2	5	Walla	Walla	1626357.76	0.43	8131788.82
3	14	Phnom	Penh	568550.77	0.96	7959710.78
4	22	myocardial	infarction	329287.5	0.77	7244324.91
5	5	Tia	Juana	1072323.8	0.1	5361619
6	18	rheumatoid	arthritis	232333.83	0.86	4182009
7	13	lactic	acidosis	306517.24	0.08	3984724.1
8	23	glomerular	filtration	146989.55	0.5	3380759.76
9	6	Pall	Mall	550344.45	0.92	3302066.68
10	14	inter	alia	221313.92	0.62	3098394.92
11	5	Deng	Xiaoping	464042.3	0.38	2320211.48
12	8	luteinizing	hormone	276643.73	1	2213149.86
13	6	Puerto	Rico	322257.59	0.42	1933545.53
14	6	intoxicating	liquors	239357.99	0.05	1436147.95
15	11	coronary	arteriography	122584.65	0.01	1348431.13
16	7	Gaza	Strip	190408.56	0.44	1332859.89
17	14	Los	Angeles	95034.73	0.89	1330486.22
18	10	Cornelius	Hackl	123255.61	0	1232556.09
19	73	et	al	16397.78	0.33	1197037.98
20	6	malacca	cane	196384.88	0.75	1178309.29
21	18	Gertrude	Stein	53185.64	0.14	957341.51
22	7	Sri	Lankan	124002.2	0.15	868015.41
23	7	Sri	Lanka	120836.19	0.76	845853.32
24	174	United	States	4779.3	0.39	831597.69
25	27	FROM	OUR	30737.4	0.16	829909.67
26	215	New	York	2956.8	0.25	635711.99
27	300	per	cent	2102.29	0.72	630686.61
28	15	Common	Pleas	41084.65	0.02	616269.82
29	8	occlusive	coronary	74912.84	0.33	599302.72
30	14	NEW	YORK	42763.07	0.09	598683.02
31	48	Creutzfeldt-Jakob	disease	12143.78	0.92	582901.64
32	6	Tel	Aviv	95668.15	0.14	574008.91
33	8	coronary	artery	70895.43	0.2	567163.46
34	17	pulmonary	artery	29598.84	0.08	503180.33
35	14	San	Francisco	33622.42	0.25	470713.85
36	12	connective	tissue	38198.43	0.67	458381.19
37	5	Monte	Carlo	91217.84	0.36	456089.2
38	6	monoclonal	antibody	64725.73	0.44	388354.37
39	12	coronary	arteries	30270.9	0.07	363250.83
40	19	Santa	Clara	17878.36	0.08	339688.89
41	8	Hong	Kong	39871.97	0.97	318975.72
42	53	Prime	Minister	5545.1	0.95	293890.14
43	7	barbed	wire	39760.82	0.74	278325.76
44	5	Rhode	Island	53102.25	0.84	265511.27
45	6	Cayman	Islands	42417.87	0.54	254507.23
46	12	Holy	Spirit	21100.72	0.34	253208.6

Table 2. Top-ranked collocations in ARCHER Late (1900-2000)

After exploring collocations, we would like to detect deeper semantic changes. This is more difficult than investigating individual collocations, which link loosely to semantics and society. The mapping from words and collocations to meaning is affected by ambiguity and synonymy: the meaning of a word heavily depends on its context, and very many words refer to broadly similar concepts. The task of semantics is thus much harder than researching single words or collocations.

However, variation between topics in a corpus can be greater than variation between time periods, or variation due to diachronic change. And the discourse of the texts themselves contains a vast amount of semantic information that can be exploited. The unit of the discourse in particular ensures that all words in a given span of text are loosely related to each other. These facts may be exploited by distributional semantics and by topic modelling, which we have discussed in sections 3.3 and 3.4.

4.2 A case study on poverty

We have chosen a case study on poverty because this has been a constant challenge for the majority of the population until recently, and it has affected the common people more than individual historical events. In Medieval England, life expectancy was only about 31 years, and besides various diseases like the plague and tuberculosis, malnutrition and famine were major causes of death. Clark (2005) shows that real wages decreased between 1500 and 1600, and only increased strongly after 1800.

Things did not immediately improve in the Modern age. Up to about 1800, life expectancy stayed below 40 years. The question of whether the industrial revolution has made England and the Western world richer or poorer is still a matter for debate. While it eventually led to increased life standards, there were long periods of hardship. Average life expectancy reached a low in the 1860s, with only 25 years in Liverpool (Szreter & Mooney 1998, C.W. 2013). Physiological data supports the observation of increased poverty: The mean height of English soldiers decreased between 1730 and 1850, only increasing afterwards owing to improving nutrition and living standards (Komlos 1998, Hatton and Bray 2010). How is poverty framed by the writers of the period? We use EEBO for the Early Modern period, and CLMET for the Late Modern period.

We first searched these two corpora for *poor* and *poverty* and synonyms extracted from the OED's historical thesaurus[6]. We found no marked increase or decrease across the periods, which on the one hand confirms that poverty prevailed and on the other hand raises the question of whether topics, associations and attitudes to poverty, the framing of poverty, has changed over time.

[6] *poor, needful, helpless, wantsome, misease, unwealy, needy, feeble, poorful, miseased, indigent, succourless, unwealthy, behove, misterous, miserable, beggarly, starved, threadbare, penurious, fortuneless, wealthless, wantful, necessitous, inopulent, egene, starveling, necessitated, inopious, destitute, want of, pauper, ruined, beggared, impoverished, pennyless, moneyless*

We analysed text spans of 10 sentences to either side of each token of the synonyms of poverty. We first used document classification to distinguish texts between the different periods. Classification accuracy was relatively low, 85% with logistic regression for two EEBO periods. The features mainly reveal linguistic changes (e.g. *unto, shall, not*), and mainly confirm changes in one topic. Early EEBO contains many religious terms (*lord, godly, pray*) as distinctive features. Also, the space of thousands of lexical features is difficult to overview. We thus turn to further data-driven approaches, in particular topic modelling and distributional semantics, which allow us to abstract from hand-driven word lists to the automatic detection of concepts, as we discussed in sections 3.3 and 3.5. We apply topic modelling to EEBO in section 4.2.1 and to CLMET in section 4.2.2.

4.2.1 A topic model of EEBO

Topic Modelling should allow us to see the topics in which poverty plays a crucial role, how contemporary writers see poverty , and which associations it evoked. We hypothesized that we would minimally find topics including *religion*, *grief and despair*, and *social and political criticism*. We compare EEBO early (1470-1599) to EEBO late (1600-1699).

In the following, we first present a run with n=8 topics. We have chosen a small number of topics for two reasons. First, we would like them to be easily interpretable. Second, while a high number of tokens allows one to identify important events and short-term trends, we wanted to seek more general trends in thought style and philosophy.

The weights, our manual label and keywords are given in Table 3. Three manual labels are presented in small caps as they do not reflect topics *per se*, but other textual differences or characteristics of topic modelling. Topic 1 (LATIN, FRENCH) gathers textual material that contains many Latin or French passages or quotes. Topic 7 is very unspecific and as it lumps together all the texts that are hard to categorize, we gave it the manual label RAGBAG.

We follow Quinn et al. (2010) and Grimmer and Stewart (2013) by using a manual verification step to assess the *semantic* and *predictive* validity (see section 3.5). As part of the semantic validity assessment, we suggest a manual label, which we give in the third column, based on the automatic keyword list, which is given in column four. We find that all topics except for topic 7 are distinctive, i.e. semantically valid.

ID	Weight	Manual Label	Keywords
0	0.39774	Prayers	god thou thy poor lord hath miserable christ man world thee life amp soul sin death mercy love gods
1	0.09261	LATIN, FRENCH	amp ye yt man good al great men thing needful god saith ben ce de time fro saint things
2	0.35528	Philosophy	men great good man make things hath reason thing world power needful nature miserable made law religion means life
3	0.14155	Bible, Church Institution	church christ god men faith people holy word true amp needful doctrine scripture ministers churches gospel book rome learned
4	0.13535	Health	water body great feeble cold make amp time drink day made eat fire blood place wine earth meat night

5	0.1558	Politics, War	king great people kings war prince time england kingdom made country majesty enemies subjects princes amp men ruined city
6	0.12269	Trade	poor money people persons time amp pay goods trade parish work made great years part year make men thereof
7	0.48246	Family, Literature, RAGBAG	poor man rich men good miserable hee house children make wife woman young time mans give gentleman money doe

Table 3. Topic Model of EEBO poverty contexts with 8 topics

Let us consider each topic individually. As part of the predictive validity we considered the most prominent or prototypical documents, i.e. at least those 5 documents which most strongly pertain to a given topic, where *mallet* reports the highest probability of belonging to the topic under investigation. It is important to read the documents that are assigned to a topic, allowing the interpreter to move between distant reading and close reading (Moretti 2013). As we chose large contexts surrounding poverty terms (10 sentences), many of the texts are not evidently related to poverty. We give excerpts of the documents below, highlighting clear poverty terms.

Topic 0 contains prayers and laudations. The meaning of *poor* in the contexts of topic 0 is our poor souls calling unto god. Example 1 is an excerpt from one of the most prominent documents for ID 0[7]:

1) for se here thy *poor* creature almost at the last gasp which calls unto the from the depth of all his *languishes* and *miseries* presenting the with his *sorrowful* & penitent soul with an humble & contrite hart the which we besech the to accept of for the love of Iohn … thy son Iesus Christ our Lord: in whose name thou hast promised to hear our prayers. (Jean de L'Espine, *The sicke-mans comfort against death and the deuill, the law and sinne, the wrath and iudgement of God*, 1590, EEBO file AD6800.xml)

Topic 1 contains many Latin and French texts, often with English comments. It is a distinct topic, even though not on purely semantic grounds. Focusing on the mainly English texts for which this topic is strongest, though, reveals a strong religious undercurrent, in which poverty is seen as a virtue, as in the text in example 2:

2) And therefore who that hath riches & loves it becomes *pour* / & they that have riches & loves *poverty* is rich (Unknown Author, *Treatise of love*, 1493, EEBO file A13930.xml)

Topic 2 contains philosophical texts. In these texts, *poor* and associated terms refer, among others, to the *conditio humana*. The beginning of the most prominent document for ID 2 is example 3:

3) Hath that Wisdom that hath made all things to operate according to their natures and provided them with whatever is necessary to that end made

[7] For ease of reading, alle examples are given in their VARDed from, i.e. after automatic spelling normalization.

myriads of noble Spirits capable of as noble operations and presently plunged them into such a condition wherein they cannot act at all according to their first and proper dispositions but shall be necessitated to the quite contrary; and have other noxious and *depraved* inclinations fatally imposed upon their pure natures? (George Rust, *Two choice and useful treatises the one, Lux orientalis, or, An enquiry into the opinion of the Eastern sages concerning the praeexistence of souls, being a key to unlock the grand mysteries of providence in relation to mans sin and misery : the other, A discourse of truth*, 1682, EEBO file A70182.xml)

Topic 3 deals with Bible discussions, exegesis and church politics. An excerpt from the most prominent document for ID 3 is example 4:

4) If any be Excommunicated without sufficient cause or by Lay Civilians to whom God never gave that power or by such Bishops or Pastors as have no just Authority for want of a true call or Consent; or if any unlawful thing be made necessary to Communion all such persons must by his own confessions hold Church-communion whether these imposers will or not; for all Christians are bound to be of some Church. (Richard Baxter, *Schism detected in both extreams, or, Two sorts of sinful separation the first part detecteth the schismatical principles of a resolver of three cases about church-communion, the second part confuteth the separation pleaded for in a book famed to be written by Mr. Raphson*, 1684, EEBO file A27028.xml)

Poverty is not an obvious central concern in topic 3. Accordingly, only one clearly poverty related term, *needful*, appears in the list of keywords in Table 3. Poor people in this topic are sinners who act against the laws of the church or commit crimes, as the sixth most prominent document illustrates in example 5:

5) the seal of Gods covenant: which is a most absurd heresy. Mark here I say a false finger for how else could the *poor* Welshman have been an absurd heretic? (Job Throckmorton, *M. Some laid open in his coulers VVherein the indifferent reader may easily see, hovve vvretchedly and loosely he hath handeled the cause against M. Penri*, 1589, EEBO file A0220.xml)

Topic 4 deals with health and medicine. Here misery comes from illness, which is often related to poverty. Example 6 is an excerpt from the most prominent document for ID 4:

6) The Pills of Rufus also are an excellent preservative against the Plague which are made after this manner following: Take Aloes and Ammoniac of each two drammes and make a composition thereof with white Wine and use the same for they are of Paulus Aeginetas description … An other preservative and very profitable for the *poor* is this that follows. Take one or two handfuls of Sorrel stepe them in a Viol in good Rose-Wine Vinegar and kepe it close … (Thomas Lodge, *A treatise of the plague containing the nature, signes, and accidents of the same, with the*

certaine and absolute cure of the feuers, botches and carbuncles that raigne in these times: and aboue all things most singular experiments and preseruatiues in the same, gathered by the obseruation of diuers worthy trauailers, and selected out of the writing of the best learned phisitians in this age, 1603, EEBO file A06182.xml)

Topic 5 deals with politics and warfare, from the present but even more frequently from the past. Military actions and praise of heroism have increased in the later period. The 17th century was marked by many political conflicts and wars, both external conflicts including the Anglo-Spanish War, the Anglo-French War, the Anglo-Dutch Wars, and internal conflicts like the English Civil War, the Irish Confederate Wars, and the Monmouth Rebellion. Important sources of the conflicts were the reformation and nationalism. While wars are often glorified in the texts, in practice they bring poverty and misery to the population, particularly in urban areas. Surprisingly from today's perspective, *poor* and synonyms usually refer to the tragic end of a war hero, much less to the sufferings of the population due to war. Example 7 is an excerpt from one of the most prominent documents for ID 5:

7) Caesars Marches were so swift that he and his Army passed the Craggy mountains of the Alps before the Roman Senate could have intelligence by their Scouts that he was departed out of France. (D. P. P., *The six secondary causes of the spinning out of this vnnaturall warre* 1644, EEBO file A54412.xml)

Example 8 illustrates the the tragic end of a hero:

8) David upon Amnon his son for the Rape of Tamer was the cause of the Murder of Amnon of the rebellion and of the *miserable* end of Absalom. (D. P. P., *The six secondary causes of the spinning out of this vnnaturall warre* 1644, EEBO file A54412.xml)

Topic 6 deals with trade and economy and contains many legal texts. Here *poverty* and related terms typically refer to the absence of money, as we can see in example 9, an excerpt from one of the most prominent documents.

9) Note That the Churchwardens and Overseers Forfeitures of the Churchwardens and Overseers for Neglect in their Office. for every Default and Negligence in their Office about the *Poor* every of them forfeits 20 s. The default to be proved either by Confession or Examination of Witnesses and is to be levied by the New Churchwardens and Overseers or one of them on Warrant by distress and sale … (Robert Gardiner, *The compleat constable directing all [brace] constables, headboroughs, tithingmen, churchwardens, overseers of the poor, surveyors of the highways, and scavengers in duty of their several offices according to the power allowed them by the laws and statutes, continued to this present time*, 1692, EEBO file A42380.xml)

Example 10 is a further excerpt from one of the most prominent documents for ID 6, this time illustrating the theme of trade, which dominates this topic:

> 10) Discharging all Merchants and Skippers or any other our Subjects to export forth of this our Kingdom any Goods or Commodities that are or shall be declared to be Staple Commodities to any other Port or Place in the Nether-lands but only to the said Staple-port and Town of Camphire in Zealand under the Pains and Certifications mentioned in the seeds Acts of Parliament and Acts of the Convention of Burghs which Pains and Penalties We ordain to be exacted from the Transgressors with all rigour and that they be further proceeded against as our Council shall find Cause. (Scotland Privy Council, *A proclamation for observing the staple-port at Camphire*, 1692, EEBO file B05662.xml)

Topic 7, finally, looks like a less distinctive topic. It is the topic with the highest weight, thus most textual mass refers to it. Checking the most prominent texts here reveals sympathy and affection at a personal level as *poor* is often used as a part of an address. Topic 7 is partly an unspecific ragbag topic. Some of the texts describe family scenes, and often come from the literary genre. 11 is an example:

> 11) Oh *miserable* Blacke-pudding if I can tell which is the way to my Masters house I am a Red-herring and no honest Gentleman. (Willian Haughton, *English-men for my money: or, A pleasant comedy, called, A woman will haue her will*, 1616, EEBO file A02800.xml)

We expected to find representations of *religion*, *grief and despair*, and *social and political criticism*. Religion appears centrally, grief and despair are also present, for example *death* in topic 0, and *miserable* in topics 2 and 7. But we did not see any clear sign of social and political criticism. The evident importance of religion is a mirror of society. The view of religion on poverty, and the glorification of national heroes partly explain the absence of social criticism and the fact that sadness and despair facing poverty are less present than expected. poverty is first and foremost a virtue, be it for religious or nationalist reasons.

Next, we consider which topics are more prominent in early EEBO (1470-1599) and which in late EEBO (1600-1699). The relative proportions are given in Figure 1. We can see that Latin and French texts, but also texts reprimanding us that poverty is a religious virtue, dominate the early period. Also prayers are slightly more frequent in the early period. Health (4), Family and Literature (7) are important in both periods. In other topic model experiments with more topics, we sometimes saw a clearer distinction between everyday or core topics and Arts and Literature, where the latter is rising. The fact that exegesis and church politics are more important than prayers may indicate a shift from devout belief to power politics. Politics and wars are generally becoming more important, and trade even more so, showing the beginning of colonization, and of early modern philosophy, which conventionally starts with René Descartes' skepticism and his publication of the *Discourse on Method* in 1637.

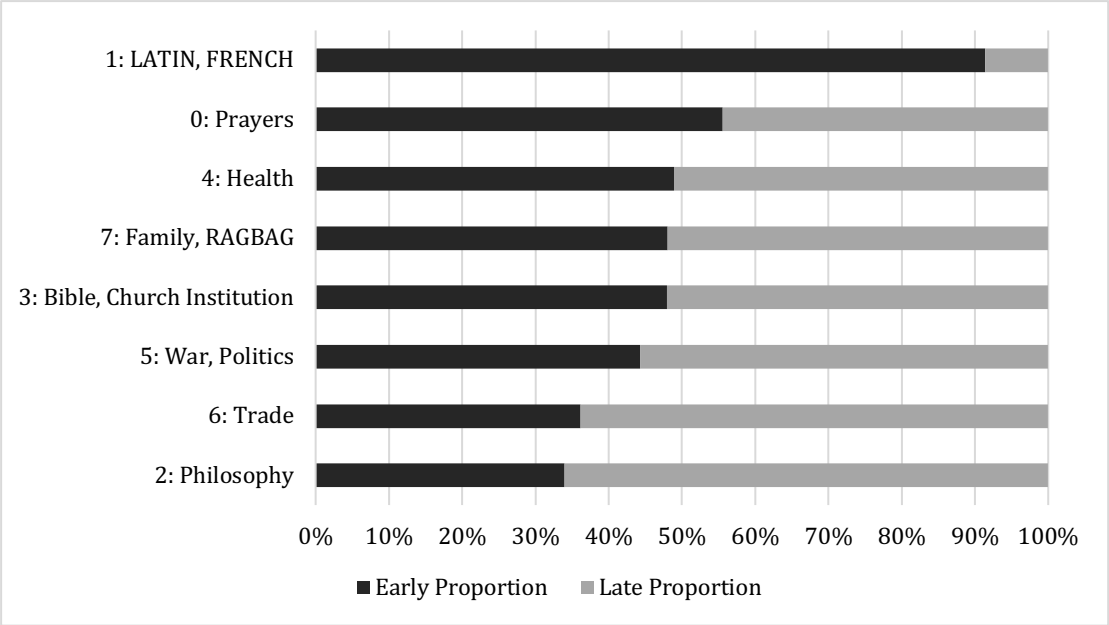

Figure 1. Relative proportion of topics, comparing Early and Late EEBO period

4.2.2 A topic model of CLMET

CLMET includes material from 1710 to 1920, split into three periods of 70 years each. We also use contexts of 10 sentences each around synonyms of *poverty*.

We set the number of topics to 20 this time, in order to get slightly finer grained semantic classes. Table 4 gives the 20 topics, with keywords in the last column. The (arbitrary) topic ID is in the first column, our interpretation in the second, and the weight of the topic in the third column.

ID	Manual Interpretation	Weight	Keywords
0	daytimes	0.2245	mr poor **day time** house **morning night** letter mrs till good home **evening** lady room london miss left long
1	religion	0.0684	**church god** world **religion** man men **christian faith religious holy st** poor people great **divine spirit** true **christ worship**
2	novel characters	0.0437	thou thy thee poor ye art ha uncle ll toby hast trim corporal man honour joseph father wilt replied
3	death & misery	0.3003	**life** poor **heart love death** mother father **miserable** day world **god** mind child long **soul** man **hope** time **thought**
4	roman heroes	0.0608	**emperor war roman empire** thousand city feeble hundred people **rome arms army troops public italy reign military** years **barbarians**
5	king & court	0.0922	**lord king sir duke** mr **queen** great **prince court house england** poor **royal english** made **parliament henry** country son
6	books & literature	0.0907	mr **book author read poet great books written genius work poetry letters works writing history write wrote** years life
7	money	0.1468	**poor** people country **money** great **labour** part **pay price** year **pounds trade** time years state **work hundred** england land
8	animals	0.1523	poor **horse** man back head made men **dog** fellow time **master** great half round **horses** foot hand tom road
9	**RAGBAG**	**0.3058**	poor man good make thing people things thought time ll day put told money made give fellow world life
10	seafaring	0.0531	**captain ship sea men boat board island shore** made **water ships found land wind vessel** english time **sail** em
11	**RAGBAG**	**0.3861**	time made found person mr received present man immediately gave manner thought occasion part house means friend conduct miserable
12	landscapes	0.1031	**country great trees water miles place small land large village** found long side **river part forest houses sun ground**
13	eat & drink	0.0726	**bread** poor **water food eat** day **good drink wine meat** people children **tea** men **dinner** made house small work
14	france	0.0280	**de paris la** poor **france** king **french** men **louis national le madame** day man ye hundred majesty **monsieur saint**
15	virtue	0.2620	man good men life **nature mind characte**r great world women **society virtue** make human pleasure reason natural sense people
16	family	0.1031	mrs poor **lady miss mother r**achel **young dear woman child** mr **father sister husband wife daughter** lord **family girl**
17	science	0.1687	**nature** man **things general fact** men great **system case knowledge** power state present form human **matter time existence** work
18	terms of address	0.2187	poor **sir lady good dear mrs mr** make man **honour** give heart **father madam** hope **miss love** great young
19	body & movement	0.2637	**eyes face** room door **hand** poor **looked** back **head** night voice round time **stood** light **hands** moment **turned sat**

Table 4. Topic Models for texts with *poverty* from the CLMET corpus. Column 2 gives our manual assessment of the automatic topics, column 3 the relative weight, and column 4 the automatic keywords

There are two topics (9 and 11) which are difficult to interpret, are very unspecific, and which have high importance (30% and 38% of the text material are assigned to them). It often happens in topic modelling that a miscellaneous or ragbag category is formed, which attracts documents that do not fit other classes well, and whose keywords are generally very frequent words, like *person*, *time*, *made*. The other 18 topics are straightforward to interpret, ensuring semantic validity, and inspecting the documents confirmed our expectations.

As we want to discover if topics and thought styles have changed over time, we now assess the contribution of each topic per period, in Figure 2 and in Table 4. We have sorted them from earliest to latest topic. We can see that the topics of *Roman heroes*, *seafaring, virtue, King & court* are clearly decreasing. The topics of *death & misery* and of *money* are most present in the middle period (1780-1850), which fits with our non-linguistic data (the height of soldiers was smallest in 1850, and life expectancy lowest in the big cities). Surprisingly, no direct reference appears to the Irish famine. The fact that Ireland exported food to England, which was affected by poor harvests and crop disease, is an important reason. Let us consider this topic in more detail.

Example 12 illustrates the most central theme in topic 3: death and the suffering of the bereaved.

> 12) She sunk upon my bosom, and expired! nor sigh nor groan gave warning of her death, she closed her eyes, and slept for ever! No words can paint the grief and distraction, of her unhappy husband … (CLMET 3_1_1_45)

A further central theme in topic 3 is art. There are many plays. In the following example 13, we see the poet's despair:

> 13) I wander without knowing where - I speak without knowing to whom, - and I look without knowing at what. - Heigho! how my *poor* heart flutters in my breast ! (CLMET 3_1_1_42)

The documents in this topic are full of horror and misery, illustrated by example 14.

> 14) Mine has been a tale of horrors; I have reached their acme, and what I must now relate can but be tedious to you. Know that, one by one, my friends were snatched away; I was left desolate. My own strength is exhausted (CLMET 3_1_2_155)

Surprisingly though, financial poverty and starvation appear as only a relatively small subtopic. The first clear example that we found when reading the most prominent document is only at rank 32. Example 15 contains the following excerpt:

> 15) We are the children of misfortune; [Agatha -] *poverty* 's chilling grasp nearly annihilates us. Our poor blind father, now the inmate of yon cottage - he who has been blessed with prosperity to be thus reduced (CLMET 3_1_2_150)

The big weight of the topic France (ID 14) partly reflects the *Napoleonic* wars.

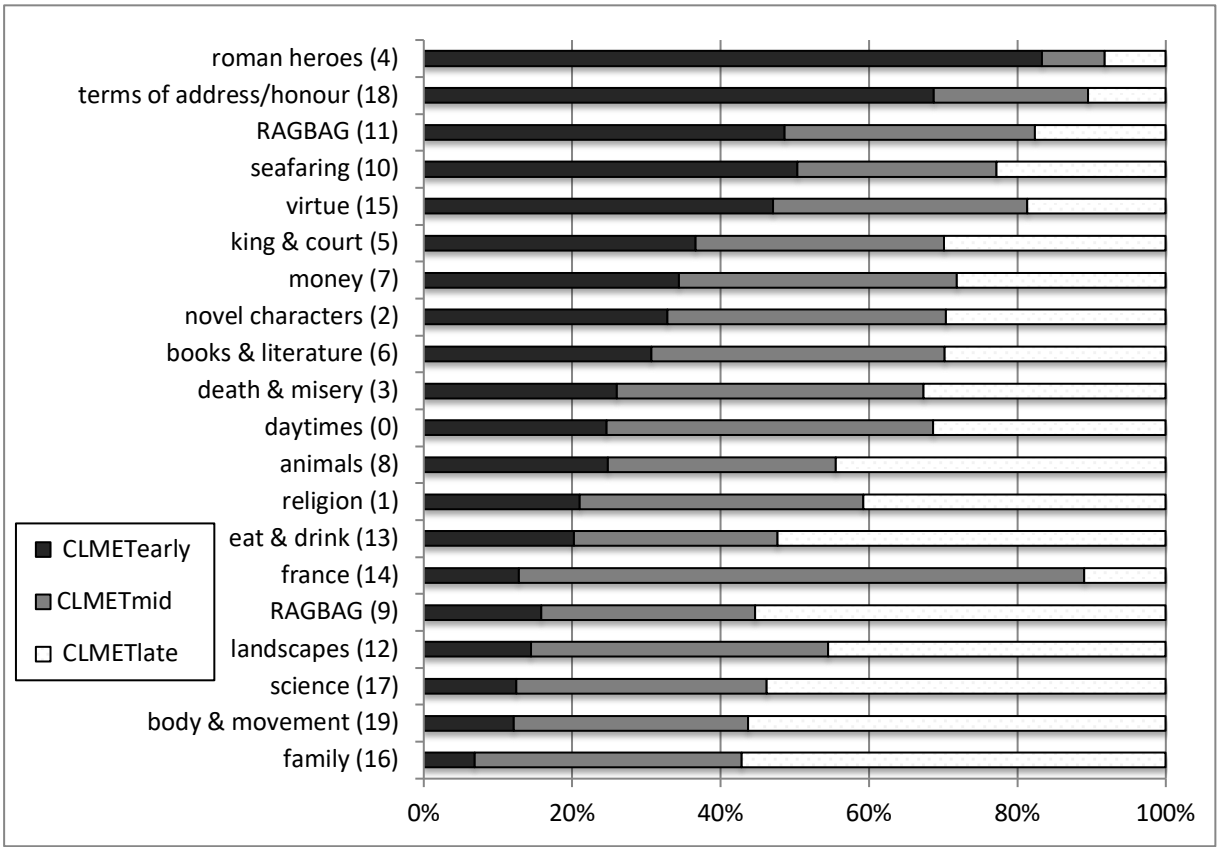

Figure 2. Distribution of the topics across the three periods of CLMET

Topic	CLMETearly (1710-1780)	CLMETmid (1780-1850)	CLMETlate (1850-1920)
roman heroes (4)	8.97%	0.92%	0.88%
terms of address/honour (18)	20.79%	6.29%	3.17%
RAGBAG (11)	15.22%	10.55%	5.53%
seafaring (10)	3.21%	1.71%	1.46%
virtue (15)	10.14%	7.39%	4.02%
king & court (5)	3.06%	2.79%	2.49%
money (7)	5.54%	6.04%	4.53%
novel characters (2)	1.73%	1.98%	1.56%
books & literature (6)	2.29%	2.94%	2.22%
death & misery (3)	6.90%	10.95%	8.65%
daytimes (0)	4.96%	8.89%	6.33%
animals (8)	2.93%	3.62%	5.24%
religion (1)	1.02%	1.85%	1.98%
eat & drink (13)	1.17%	1.59%	3.03%
france (14)	0.56%	3.34%	0.48%
RAGBAG (9)	4.13%	7.53%	14.45%
landscapes (12)	1.56%	4.32%	4.91%
science (17)	2.03%	5.50%	8.78%
body & movement (19)	3.07%	7.98%	14.24%
family (16)	0.72%	3.80%	6.05%
∑	100.00%	100.00%	100.00%

Table 4. Numerical data for Figure 2. Distribution of the topics across the three periods of CLMET

The fact that *religion* is most important in the late period seems surprising after the constant decline in EEBO. It is due to the corpus collection strategy used for the CLMET corpus, which does not contain texts from the religious genre. This also explains its generally low importance of only 1-2% (see Table 4).

We first consider a selection of topics that has decreased, then we turn to increasing topics. One of the topics to decrease most is ID 18 *honour & terms of address*. An excerpt from the most prominent document, i.e. the document belonging most strongly to this topic, is example 16:

16) alas! sir, it is out of your power to preserve my poor girl. -- O my child! my child! she is undone, she is ruined for ever! " `` I hope, madam, " said Jones, `` no villain " -- `` O Mr Jones! " said she, `` that villain who yesterday left my lodgings, hath betrayed my *poor* girl; hath destroyed her. -- I know you are a man of honour. You have a good -- a noble heart, Mr Jones. (CLMET 3_1_1_23)

A further topic to have decreased strongly is ID 15, *virtue*. The most prototypical document is excerpted in example 17:

17) Where we expect a beauty, the disappointment gives an uneasy sensation, and produces a real deformity. An abjectness of character, likewise, is disgustful and contemptible in another view. Where a man has no sense of value in himself, we are not likely to have any higher esteem of him. And if the same person, who crouches to his superiors, is insolent to his inferiors (as often happens), this contrariety of behaviour, instead of correcting the former vice, aggravates it extremely by the addition of a vice still more odious. (CLMET_3_1_1_30)

Helping the poor is not a frequently mentioned virtue, but charity is pointed out as a virtue in the 10th most prominent document in example 18:

18) I have, in truth, observed, and shall never have a better opportunity than at present to communicate my observation, that the world are in general divided into two opinions concerning charity, which are the very reverse of each other. One party seems to hold, that all acts of this kind are to be esteemed as voluntary gifts, and, however little you give (if indeed no more than your good wishes), you acquire a great degree of merit in so doing. Others, on the contrary, appear to be as firmly persuaded, that beneficence is a positive duty, and that whenever the rich fall greatly short of their ability in relieving the *distresses* of the *poor*, their pitiful largesses are so far from being meritorious, that they have only performed their duty by halves, and are in some sense more contemptible than those who have entirely neglected it. To reconcile these different opinions is not in my power. I shall only add, that the givers are generally of the former sentiment, and the receivers are almost universally inclined to the latter. (CLMET 3_1_1_23)

Charity seems to be the only virtue that is needed to address the problem of poverty – we did not find other direct connections nor any social criticism. Some discussions on virtue stress the importance of education, though, where a firm hand is recommended, as illustrated in example 19.

19) To give an example of order, the soul of virtue, some austerity of behaviour must be adopted, scarcely to be expected from a being who, from its infancy, has been made the weathercock of its own sensations. Whoever rationally means to be useful, must have a plan of conduct ; and, in the discharge of the simplest duty, we are often obliged to act contrary to the present impulse of tenderness or compassion. Severity is frequently the most certain, as well as the most sublime proof of affection; and the want of this power over the feelings, and of that lofty, dignified affection, which makes a person prefer the future good of the beloved object to a present gratification, is the reason why so many fond mothers spoil their children, and has made it questionable, whether negligence or indulgence is most hurtful: but I am inclined to think , that the latter has done most harm. Mankind seem to agree, that children should be left under the management of women during their childhood. Now, from all the observation that I have been able to make, women of sensibility are the most unfit for this task, because they will infallibly,

carried away by their feelings, spoil a child 's temper. (CLMET 3_1_2_107)

We now turn to topics that have increased strongly. The latest CLMET period continues and accelerates changes that the middle period started: we expected the rapid progress of science, which is equally an effect of and a reason for industrialisation. The increase of the topic *family* reflects an increase in descriptions of affections and feelings. That we find more descriptions of *landscapes*, *body & movement* may be more surprising.

Topic 12 (*landscapes*) on the one hand contains travel descriptions and logs, often including the poor living conditions of the native people. These descriptions are in line with discovering exotic lands and the new world, and the increasing importance of trade, as illustrated in example 20.

20) The banks at first were low and marshy and intersected by numerous channels; the principal tree was a long, coarse-leaved palm, and there were great beds of wild cane and grass, amongst which we occasionally saw curious green lizards … As we proceeded up the river, the banks gradually became higher and drier, and we passed some small plantations of bananas and plantains made in clearings in the forest, which now consisted of a great variety of dicotyledonous trees with many tall, graceful palms; the undergrowth being ferns, small palms, Melastomae , Heliconiae , etc. . The houses at the plantations were mostly *miserable* thatched huts with scarcely any furniture (CLMET 3_1_3_223)

It is noticeable, particularly in the later period, that many nature descriptions extend beyond a botanical or geographical interest and stress the pure beauty of nature (example 21).

21) when the sun had set, we beheld immense mountains and precipices overhanging us on every side, and heard the sound of the river raging among rocks, and the dashing of water-falls around. The next day we pursued our journey upon mules; and as we ascended still higher, the valley assumed a more magnificent and astonishing character. Ruined castles hanging on the precipices of piny mountains; the impetuous Arve, and cottages every here and there peeping forth from among the trees, formed a scene of singular beauty. (CLMET 3_1_2_155)

In literature, the new style of *literary realism* is evident, which we will also see in the next subsection. Topic 19 (*body & movement*) might also be seen in this light as an artistic interest in the body, as illustrated in example 22..

22) He turned his canvas towards them, and upon it was a picture of Ethel, standing beside the sea, surrounded by a band of fairies, who pulled her skirts and tried to make her look at them; but her head was turned away . `` *Poor* Ethel, " said Rachael; `` how they pull. " (CLMET 3_1_3_281)

4.3 Charles Dickens' social criticism

One aspect of poverty that we had expected to find in EEBO and CLMET was social criticism, but we largely failed. We ran topic models with more topics, and models combining EEBO and CLMET. The combined models have the disadvantage that, because EEBO is a random collection of published books, while CLMET is a sampled corpus, genre differences and changes in topics are equally reported. For example, religion is underrepresented in CLMET, while the genre of personal letters, which is absent in EEBO, creates new apparently recent topics. But combined models with a high number of topics show very little social criticism. For example, with 50 topics on EEBO combined with CLMET, we obtained only one topic that seems to express social criticism. Its keywords are *system, class, fact, population, work, general, society, present, state, pg, time, found, classes, large, great, made, case*.

Inspecting the prototypical documents revealed passages like the following in examples 23 and 24.

23) It is from the class of females above described, that we naturally look for the highest tone of moral feeling, because they are at the same time removed from the pressing necessities of absolute poverty, and admitted to the intellectual privileges of the great; and thus, while they enjoy every facility in the way of acquiring knowledge, it is their still higher privilege not to be exempt from the domestic duties which call forth the best energies of the female character.

24) If anything be urgently wanted, it is a plan for preventing the growth of the criminal class; and this probably is not so difficult as it may appear. Of course, till there be a far broader system of public education than now prevails, the criminal population will never want recruits. Nevertheless, even with our present imperfect educational arrangements, something might be done. The criminal class is discovered to be on the whole a narrow class. The practice of living by depredation runs in families, and clings to individuals. The police of any given town could put their hand on almost every person who lives by fraud, theft, and robbery. They could at a day's notice secure nearly every one of them.

We realize that visions of a society that can overcome poverty were rare in the period. Searching further, with mixed methods, we finally found three texts by Charles Dickens that are part of CLMET. It may be that he was an outlier in terms of his vision of society. We investigate this question further in the following section.

4.3.1 Frequency of poverty in Dickens

We investigate the associations that Dickens' writings have to poverty, to assess whether the claims that have been made concerning his writing can be substantiated with data-driven methods. His writings come from the period from about 1840 to 1870, a period in which smaller body height and low life expectancy, both partly caused by malnutrition, starvation and poverty, were particularly acute (Komlos 1998,

Hatton and Bray 2010, C.W. 2013). Kailash (2012) and Mahlberg (2013) suggest that social concerns are reflected in his writings.

> Charles Dickens was one of the most important social critics who used fiction effectively to criticize economic, social and moral abuses in the Victorian era. He showed **compassion and empathy towards the vulnerable and disadvantaged segments of English society**, and contributed to several important social reforms (Kailash 2012:1, emphasis added)

> By linking people to things and minds to machines, Dickens's techniques of characterisation can be read as addressing the relationship between industrial mechanisation and society (Mahlberg 2013: 30)

We collected a small corpus of his works, consisting of 8 of his popular novels, where we expected 7 to contain the topic of poverty centrally, and one less (*Great Expectations*), in total 1.6 million words:

- A Christmas Carol
- David Copperfield
- Great Expectations
- Hard Times
- Nicholas Nickleby
- Oliver Twist
- The Pickwick Papers
- A Tale of Two Cities

We compare the collection to CLMET. The middle and late CLMET periods are the times of the publications of Dickens' works. Table 5 compares frequencies. As expected, *poor* and *poverty* are important concepts in Dickens writings (though less so in *Great Expectations*). We can also see that the descriptive adjective *poor* seems to be used more than the abstract concept of *poverty*.

poor	*poverty*	Text Source
928.045	**61.869**	Dickens Christmas Carol
522.568	*19.561*	Dickens David Copperfield
410.830	*5.335*	Dickens Great Expectations
545.872	*9.411*	Dickens Hard Times
805.321	**55.327**	Dickens Nicholas Nickleby
313.292	36.275	Dickens Pickwick Papers
626.426	**57.602**	Dickens Tale Of Two Cities
614.872	31.054	Dickens_Oliver Twist
393.178	45.524	CLMET 3.1 early
471.119	36.285	CLMET 3.1 mid
414.520	32.100	CLMET 3.1 late

Table 5. Frequency (per million words) of *poor* and *poverty* in Charles Dickens and CLMET. **Boldprint** indicates high frequency, *italics* indicate low frequency, the darkness of the shading reflects the deviation from the mean.

4.3.2 A topic model of Dickens works

As the concept of poverty seems to be important in our selected Dickens novels, we ran a topic model on them[8]. To make sure that the topic model does not mainly follow individual novels, we replaced all proper names by the placeholder 'np'.

The raw output, without manual interpretation, is given in Table 6. The topics which are easily interpretable are presented in bold print. We can see topics describing *horses and carts*, allegedly *honourable company managers*, poor as a term of *affection in the family*, the *copyright* line, descriptions of *body parts*, *light & dark*, *vernacular speech*, *writing*, *descriptions of rooms* and scenes of *drinking and eating*. The last two in particular, but also affection in the family, follow the new style of literary realism, describing everyday scenes of the working classes. Descriptions of body parts were also detected by Mahlberg (2013: 100-127,155) using lexical bundles. We think that topic models are a useful addition to her stylistic investigation.

ID	Importance	Keywords
0	0.48126	great time lady young gentleman mind state friend person made conversation moment part appeared object place
1	**0.0952**	**coach scrooge horse uncle horses carriage chaise guard road cart box stopped coachman door baron roads passengers**
2	0.28014	man cried back head boy men hands round dog blood moment made dead hold arms cry hand body feet
3	**0.12916**	**gentlemen company people ladies great gentleman public men man friends honourable manager general party stage**
4	**0.01806**	**work works electronic terms agreement donations copyright copy full access paragraph trademark laws set fee**
5	**0.44488**	**heart father life child love poor day mother hope world knew young thought tears years long mind dear brother**
6	**0.59361**	**face eyes hand looked head sat hands man chair back turned side round stood time put arm voice made**
7	0.20197	coat man hat black hair gentleman head white red large half great boots green small eye round pair legs
8	0.79153	np dear don good aunt returned make head asked ll mother cried thought thing mind suppose pretty time poor
9	**0.20018**	**light night place air wind water great people dark sun sea cold stone passed streets long men high lay**
10	**0.06381**	**wery em ain gen replied afore ll wos wot father ere fur sir ve thou good man vith inquired**
11	**0.13364**	**read book paper letter pocket put office clerk papers pen gentlemen prisoner business case writing court desk wrote**
12	0.27174	np sir man replied gentleman ma lady stranger beg returned ll don speak hear pardon call hope business rejoined
13	1.14976	np replied friend inquired friends observed give exclaimed end master returned long smile cried hat called whisper
14	**0.43416**	**time night day home morning house back long found evening place made thought left good hour days half bed**

[8] We use the entire collection of 8 Dickens novels, not only contexts containing poverty.

15	0.38283	made family great good manner business life fact present man make opinion short question confidence sister thing
16	**0.2931**	**door room house back window open street light stairs opened bed shut candle fire night walked upstairs dark looked**
17	**0.14005**	**glass water table wine dinner tea bottle drink fire bread hot good put drank half waiter cold drinking punch**
18	**0.13506**	**money madame defarge twenty hundred man pounds time year wife years good business thousand pound shop day pay**
19	0.2213	boy gentleman young replied lady man dear fat ll inquired jew doctor half boys sir em good girl woman

Table 6. Topics arising with Topic Modelling from the selected Dickens novels

When interacting with the texts we discovered that topic 3 contains scathing criticism and irony. There are typically no direct comments, no abstract discussions of poverty, but a very descriptive and suggestive style. Think of the famous beginning of *Oliver Twist*:

> Although I am not disposed to maintain that the being born in a workhouse, is in itself the most fortunate and enviable circumstance that can possibly befall a human being, I do mean to say that in this particular instance, it was the best thing for Oliver Twist that could by possibility have occurred. The fact is, that there was considerable difficulty in inducing Oliver to take upon himself the office of respiration,– a troublesome practice, but one which custom has rendered necessary to our easy existence; and for some time he lay gasping on a little flock mattress, rather unequally poised between this world and the next: the balance being decidedly in favour of the latter. Now, **if, during this brief period, Oliver had been surrounded by careful grandmothers, anxious aunts, experienced nurses, and doctors of profound wisdom, he would most inevitably and indubitably have been killed in no time. There being nobody by, however, but a pauper old woman, who was rendered rather misty by an unwonted allowance of beer; and a parish surgeon who did such matters by contract**; Oliver and Nature fought out the point between them. The result was, that, after a few struggles, Oliver breathed, sneezed, and proceeded to **advertise to the inmates of the workhouse the fact of a new burden having been imposed upon the parish**, by setting up as loud a cry as could reasonably have been expected from a male infant who had not been possessed of that very useful appendage, a voice, for a much longer space of time than three minutes and a quarter. (Charles Dickens 1838. Oliver Twist, London: Bentley. Emphasis added)

4.3.3 Synonyms and associations

So what does poverty mean for Dickens? Does he really present a different view from his contemporaries? Let us give a computational answer, by computing a distributional semantic model trained on our selected works by Dickens. We have used the *R* library *WordVectors*[9]. We used the library to query for synonyms of *poverty*. As mentioned, distributional semantics is also likely to report associations, or

[9] https://github.com/bmschmidt/wordVectors

even antonyms, particularly in ironic writing. We can then compare Dickens' synonyms with those of the corpus from the period, i.e. CLMET mid and late period. The synonyms of *poverty* are given in Table 7. While his contemporaries (in CLMET Period 2 and 3) mainly associate poverty with misery and disgust, Dickens thinks of its wrongs and oppressions, and the grotesque debauchery and cupidity at the opposite end of the social strata.

The fact that the lists of synonyms show related terms and associations indicates that semantic validity is high. That they are largely different shows us that predictive validity is also high.

Distr. Sem.	**bold** = empathy & social criticism		*italics*=disgust & misery		
SYNONYM	words	**1780-1850**	words	**1850-1920**	words
Dickens	1619929	**CLEMT p2**	1640497	**CLMET p3**	1535183
	word	**sim to poverty**	**word**	**sim to poverty**	**word**	**sim to poverty**
1	poverty	1	**1** poverty	1	**1** poverty	1
2	**debauchery**	0.5651	**2** debasing	0.5461	**2** degradation	0.5695
3	**wrongs**	0.5636	**3** *misery*	0.5338	**3** *destitution*	0.5530
4	**cupidity**	0.5542	**4** cravings	0.5214	**4** *miseries*	0.5468
5	breasts	0.5442	**5** *violating*	0.5152	**5** *dregs*	0.5398
6	wealth	0.5413	**6** indigence	0.5092	**6** alleviate	0.5265
7	**oppression**	0.5365	**7** *punishments*	0.5033	**7** compensations	0.5220
8	**sickness**	0.5335	**8** debase	0.4981	**8** *squalid*	0.5176
9	riches	0.5302	**9** hardens	0.4974	**9** *misery*	0.5104
10	unrelenting	0.5268	**10** *untaught*	0.4946	**10** penury	0.5019
11	joys	0.5214	**11** degradation	0.4936	**11** *squalor*	0.4984
12	**griefs**	0.5176	**12** immoderate	0.4760	**12** **commiseration**	0.4903
13	**hardship**	0.5168	**13** unassisted	0.4756	**13** privations	0.4876
14	baseness	0.5152	**14** automaton	0.4745	**14** brotherhood	0.4855
15	privation	0.5132	**15** luxury	0.4723	**15** **sufferings**	0.4851
16	*barbarous*	0.5130	**16** extravagance	0.4713	**16** *lice*	0.4811
17	*destitute*	0.5102	**17** tutors	0.4689	**17** toil	0.4800
18	**heartless**	0.5081	**18** **profligacy**	0.4685	**18** *intoxication*	0.4794
19	*sordid*	0.5050	**19** *wretchedness*	0.4675	**19** *thriftless*	0.4794
20	**purest**	0.5030	**20** *destitution*	0.4667	**20** hovels	0.4791

Table 7. Synonyms of *poverty* in Dickens, compared to contemporary periods of CLMET

In the light of Table 7, Charles Dickens' works seem all the more visionary, as his contemporaries and even later generations largely had different views.

5 Conclusions

We have investigated how data-driven approaches can systematically detect changes, in society, history, thought styles and language. We have critically assessed data-driven methods, at the level of words (multi-word units, collocations), topics (topic modelling) and semantics (distributional semantics), and presented two related case studies.

We have set out to address (research question 1) the broad research question of whether the data-driven methods manage to detect meaningful, interpretable and consistent patterns. The answer to that research question needs to follow from our two interrelated pilot studies: can patterns, topics and associations of poverty be detected

(research question 2)? How did the contemporary writers, from 1470 to 1910, frame poverty (research question 3)? Were Dickens' views really different, visionary, and are they reflected in the data and detected by corpus-driven approaches (research question 4)?

In response to research question 2, patterns change strongly across periods. The list of collocations has hardly any overlap and many topics suggested by topic modelling increase and decrease over time. In answer to research question 1, most of the topics can be interpreted meaningfully. We can see how collocations reflect trends of the periods, how today's world has fragmented into specialized technical fields, and how the role of religion has diminished. The validation of topic models with semantic and predictive evaluation showed that most topics are internally coherent and discriminate well among other topics, although the method produces one or two unspecific ragbag topics.

Concerning research question 3, we note that poverty was, first, rarely framed as a social problem, more as a religious virtue, and second as a feature of the low, untaught, and criminal classes. The reaction to the poverty of the low classes was at best charity, but more often disgust and a call for law and order and more severe punishment. The first obyervation, poverty as a religious virtue, is particularly prominent in the early phase, while the second observation, the call for law and order, is particularly prominent in the late phase.

The voice of Dickens, which we investigate in research question 4, stands out as visionary and different from the prevailing opinions of his contemporaries. Our investigation of collocations has shown how much the world has changed over the centuries. Attitudes and reactions to poverty may have changed less, though. Religious indoctrination, and law and order policies remain with us today.

As further insight, we think that our investigation of topic models and associations by means of distributional semantics usefully complements Mahlberg's (2013) stylistic investigation of Dickens, delivering more topics, and adding a more content-based perspective.

In analogy to *systems biology*, our suggested approach could be called *systems history* and *systems diachronic linguistics.* But our study has several limitations and is not free from bias. First and foremost, although we follow Grimmer and Stewart (2013) in validating topics and collocations, the evaluation has to stay partly impressionistic. It is almost impossible to assess what we have missed. While we find both semantic and predictive validity convincing, both of them express precision, while recall is harder to assess. The human interpreter sees how many of the found collocations, topics or synonyms and associations are correct. On the side of recall, i.e. how many of the total topics, collocations and associated term are missed, it is much harder to make a judgement. While data-driven methods give us an opportunity to detect new patterns, they cannot tell us how many we are still missing.

In future research, we would like to improve the evaluation: which trends and events do we expect to see mirrored in the data? This would provide a first assessment of the recall of our methods. Second, we would like to incorporate into our research the lexical bundles used in Mahlberg (2013), who also concludes that mixed methods

should be used in research. Third, we would like to apply our methods to further corpora, such as large subsets of the Gutenberg Archive. Fourth, we would like to include proper names, for character studies in literary settings and describing relations and networks between important political actors.

References

Ananiadou, Sophia, Kell, Douglas B., and Tsujii, Jun-ichi. 2006. Text mining and its potential applications in systems biology. *Trends in Biotechnology*, 24, 12, 571–579.

Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". *Wired Magazine* 06.2008. Available at https://www.wired.com/2008/06/pb-theory (accessed 20.02.2020)

Aarts, Bas. 2019. "Syntactic argumentation". In Bas Aarts, Jill Bowie and Gergana Popova (eds.) *The Oxford Handbook of English Grammar*. Oxford: Oxford University Press.

Baron, Alistair and Rayson, Paul. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham, UK, 22 May 2008.

Baroni, Marco and Lenci, Alessandro. 2010. "Distributional Memory: A general framework for corpus-based semantics". *Computational Linguistics*, 36, 4, 673-721.

Bartsch, Sabine and Evert, Stefan. 2014. Towards a Firthian notion of collocation. *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern*, 2/2014, 48--61.

Biber, Douglas, Finegan, Edward, and Atkinson, Dwight. 1994. ARCHER and its challenges: Compiling and exploring A Representative Corpus of Historical English Registers. *Creating and using English language corpora, Papers from the 14th International Conference on English Language Research on Computerized Corpora*, Zurich, 1-13.

Blei. David. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

Bybee, Joan. 2007. *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press.

Church, Kenneth. 2000. Empirical Estimates of Adaptation: The chance of Two Noriegas is closer to p/2 than p^2. *Proceedings of the 17th conference on Computational linguistics*, 180–186.

Clark, Gregory. 2005. The condition of the working class in England, 1209–2004. *Journal of Political Economy*, 113(6), 1307-1340.

C.W. 2013. Did living standards improve during the Industrial Revolution? *The Economist*, September 13, 2013. <https://www.economist.com/free-exchange/2013/09/13/did-living-standards-improve-during-the-industrial-revolution> (30 December 2018)

Entman, Robert M.. 1993. "Framing: toward clarification of a fractured paradigm". *Journal of Communication*, 43(4), 51-58.

Evert, Stefan. 2006. How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik*, 177 - 190.

Evert, Stefan. 2009. "Corpora and collocations". *Corpus Linguistics. An International Handbook*, article 58, 1212-1248.

Firth, John Rupert. 1957. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*, 1-32.

Fitzmaurice, Susan, Robinson, Justyna A.., Alexander, Marc, Hine, Iona C., Mehl, Seth, and Dallachy, Fraser. 2017. Linguistic DNA: Investigating Conceptual Change in Early Modern English Discourse, *Studia Neophilologica* 89:sup1, 21-38.

Glynn, Dylan. 2010. Corpus-driven Cognitive Semantics. Introduction to the field. *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches*. Berlin, Boston: De Gruyter Mouton, 1-42.

Gries, Stefan Th.. 2010. Corpus linguistics and theoretical linguistics: a love-hate relationship? Not necessarily … *International Journal of Corpus Linguistics* 15(3). 327-343. <http://www.linguistics.ucsb.edu/faculty/stgries/research/2010_STG_CorpLingLingTheory_IJCL.pdf>

Grimmer, Justin and Stewart, Brandon. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

Grover, Claire & Tobin, Richard. 2006. Rule-Based Chunking and Reusability. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy. 873–878.

Hatton, Timothy J. and Bray, Bernice E. 2010. Long run trends in the heights of European men, 19th–20th centuries. *Economics & Human Biology*, 8(3), 405-413. <https://www.sciencedirect.com/science/article/pii/S1570677X10000225?via%3Dihub> (30 December 2018)

Hilpert, Martin, and Gries, Stefan. 2016. Quantitative approaches to diachronic corpus linguistics. Merja Kytö & Paivi Pahta (eds.), *The Cambridge Handbook of English Historical Linguistics*. Cambridge: Cambridge University Press, 36-53.

Janda, Linda A. 2013. *Cognitive Linguistics: the Quantitative Turn*. Berlin: Mouton de Gruyter,.

Jurafsky, Daniel, and Martin, James H. 2009. *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics*. 2nd edn. Upper Saddle River, NJ: Prentice-Hall.

Kailash, Sudha. 2012. Charles Dickens as a social Critic. *International Journal of Research in Economics & Social Sciences*, 2, 8, 1-51.

Komlos, John. 1998. Shrinking in a Growing Economy? The Mystery of Physical Stature during the Industrial Revolution. *Journal of Economic History*, 58, 779-802. <https://pdfs.semanticscholar.org/707f/a7f3faea996c241ab96dfa774244d23221b8.pdf> (30 December 2018)

Leech, Geoffrey, Hundt, Marianne, Mair, Christian & Smith, Nicholas. 2009. *Change in Contemporary English. A Grammatical Study*. Cambridge: Cambridge University Press.

Mahlberg, Michaela. 2015. "Literary Style". In D. Biber and R. Reppen (eds.). *The Cambridge Handbook of Corpus Linguistics*, 346-361. Cambridge: Cambridge University Press.

Mahlberg, Michaela. 2013. *Corpus Stylistics and Dickens's Fiction*. Routladge Advances in Corpus Linguistics Series, 14. New York and London: Routledge.

Michel, Jean-Baptiste, Shen, Yuan Kui, Aiden, Aviva P., Veres, Adrian, Gray, Matthew K., , Pickett, Joseph P., Hoiberg, Dale, Clancy, Dan, Norvig, Peter, Orwant, Jon, Pinker, Steven, Nowak, Martin A., and Aiden, Erez Lieberman. 2010. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 331 (6014), 176-182.

Miller, George A., Charles, Walter G. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1), 1–28.

Moreton, Emma and Culy, Chris. 2020. New Methods, Old Data: Using Digital Technologies to Explore Nineteenth Century Letter Writing Practices. In Carolin Tagg and Mel Evans (eds.). *Message and Medium: English Language Practices Across Old and New Media.* Berlin/Boston: De Gruyter Mouton.

Moretti, Franco. 2013. *Distant Reading*. London: Verso.

Oakes, Michael P. 2014. *Literary detective work on the computer*. Amsterdam & Philadelphia, PA: Benjamins.

Pecina, Pavel. 2009. *Lexical Association Measures: Collocation Extraction*. Studies in Computational and Theoretical Linguistics 4. Prague: Institute of Formal and Applied Linguistics, Charles University in Prague.

Piao, Scott, Dallachy, Fraser, Baron, Alistair, Demmen, Jane, Wattam, Steve, Durkin, Philip, McCracken, James, Rayson, Paul, Alexander, Marc. 2017. A Time-Sensitive Historical Thesaurus-Based Semantic Tagger for Deep Semantic Annotation. *Computer Speech & Language* (46), 113-135.

Rayson, Paul. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics*. 13:4, 519-549.

Sahlgren, Magnus. 2006. *The Word-Space Model: Using distributional Analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* Ph.D. thesis, Stockholm University.

Schneider, Gerold. 2014. *Applying Computational Linguistics and Language Models: From Descriptive Linguistics to Text Mining and Psycholinguistics*. Cumulative Habilitation, University of Zurich, Faculty of Arts.

Schneider, Gerold. 2020. Spelling Normalisation of Late Modern English: Comparison and Combination of VARD and Character-based Statistical Machine Translation. In Merja Kytö and Erik Smitterberg (eds.): *Late Modern English: Novel encounters*. Studies in Language Series. Amsterdam: Benjamins.

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics* 24 (1), 97-123.

Schwartz, H. Andrew and Ungar, Lyle H.. 2015. Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods. *The ANNALS of the American Academy of Political and Social Science*, 659, 1, 78–94.

Szreter, Simon, and Graham Mooney. 1998. Urbanization, Mortality, and the Standard of Living Debate: New Estimates of the Expectation of Life at Birth in Nineteenth-Century British Cities. *The Economic History Review*, vol. 51(1), 84–112. <http://www.jstor.org/stable/2599693> (1 January 2022)

Taavitsainen, Irma and Gerold Schneider. 2019. Scholastic argumentation in Early English medical writing and its afterlife: new corpus evidence. In Carla Suhr, Terttu Nevalainen & Irma Taavitsainen, eds. *From data to evidence in English language research*. Language and Computers, Volume 83. Leiden: Brill.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: Benjamins.

Turney, Peter D. and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37: 141-188.

Xie, Yu. 2011. Values and Limitations of Statistical Models. *Research in Social Stratification and Mobility*, 29, 343–349.

Yang Li-gong, Jian, Zhu and Shi-ping, Tang. 2013. Keywords extraction based on text classification. *Advanced Materials Research* 765-767. 1604-1609.