Machines in the media: semantic change in the lexicon of mechanization in 19th-century British newspapers

Nilo Pedrazzini
The Alan Turing Institute (UK)
npedrazzini@turing.ac.uk

Barbara McGillivray
King's College London (UK)
The Alan Turing Institute (UK)
barbara.mcgillivray@kcl.ac.uk

Abstract

The industrialization process associated with the so-called Industrial Revolution in 19th-century Great Britain was a time of profound changes, including in the English lexicon. An important yet understudied phenomenon is the semantic shift in the lexicon of mechanisation. In this paper we present the first large-scale analysis of terms related to mechanization over the course of the 19th century in English. We draw on a corpus of historical British newspapers comprising 4.6 billion tokens and train historical word embedding models. We test existing semantic change detection techniques and analyse the results in light of previous historical linguistic scholarship.

1 Introduction

Started in the 18th century in Great Britain, the industrial mechanization saw a dramatic acceleration in the 19th century. New machines were introduced in different industries at a rapid pace and the ever more pervasive automation of manufacture meant large-scale reorganization and movement of the workforce throughout the territory. This had profound repercussions on many aspects of daily life from a cultural, political, and social perspective.

The English language, particularly its lexicon, used by 19th-century sources to describe these changes reflected the same rapid pace at which the objects and the societal landscape had been shifting, making it an important, and yet understudied, research topic. Previous studies on the English language in the 19th century have focussed on how the changes observed in the lexicon of the period often reflect 'new interpretations given to older words in a time of changing societal values' in Victorian Britain (Görlach, 1999, 132), as in the shift in the usage of words describing men and women (Bäcklund, 2006), or have highlighted the plethora of neologism and new loanwords introduced as a result of the Industrial Revolution (Kay and Allan, 2015, 20; Bergs and Brinton, 2012). As Görlach (1999, 133) also notes, meaning change (besides mere new word formations) 'is best illustrated from semantic fields relating to the new technologies that rapidly became part of everyday experience, such as the field of vehicles/transport/traffic'.

In this paper we investigate the issue of tracing these subtle shifts at scale using computational methods. Drawing from examples of lexical semantic change in 19th-century English from previous literature, we train diachronic word embedding models on a very large collection (4.6 billion tokens) of digitized 19th-century British newspaper articles. We then compare these data-driven analyses with previous qualitative studies, to verify the extent to which historical language models reflect expert knowledge. In addition to validating the computational models, we assess how these methods can be employed to answer new complex questions on the linguistic effects of mechanization and other historical events.

Using historical newspapers as a data source presents specific methodological challenges, and in particular historical (issues of representativeness, Beelen et al. 2022) and computational (processing OCR'd collections, van Strien et al. 2020) complexities. However, given the size of newspaper archives and the possibility to sample them by variables of interest (e.g. time period, political leaning, place of circulation or publication), these sources are a very good fit for large-scale analyses of lexical change in periods of on-going deep societal changes. This is also shown by the growing number of projects which use historical newspapers as sources for large-scale semantic processing and data-driven historical analysis, including News-

Eye,[1] Translantis,[2] Impresso,[3] and Living with Machines.[4]

This work is the first to provide a large-scale analysis of the English lexicon of mechanisation in the 19th century. From a methodological point of view, our dataset presents challenges that are shared by other historical newspaper archives and thus our research can inform similar studies on other languages. From the point of view of historical linguistics and historical research, we present the first study of the English lexicon of mechanisation that employs computational techniques, which allows us to compare automatically detected semantic changes with those identified by close-reading methods in previous literature.

2 Previous work

According to Görlach (1999) and Mugglestone (2008), the 19th century was a pivotal period in the history of English, when its lexicon underwent a significant transformation in both spoken and written sources, although the academic literature has paid less attention to Late Modern English (1700-1950) compared to other periods in the history of the English language (Kytö et al., 2006). In recent years a number of NLP studies have proposed algorithms for the automatic detection of lexical semantic change from historical texts using word type and token embeddings (Hamilton et al., 2016; Tsakalidis et al., 2019). Algorithms based on type embeddings have been shown to perform best in the 2020 SemEval shared task (Schlechtweg et al., 2020) and they typically consist of the following steps: the corpus of interest is divided into time-dependent slices; then word embedding models are trained from each subcorpus and their spaces aligned. Finally, the cosine similarity between a word's embedding in the first (or last) space and its embedding in each of the spaces is computed. If the similarity is below a predefined threshold (i.e. the two embeddings are sufficiently different), the word is marked as a potential candidate for semantic change. In few cases these algorithms have been applied in real-world digital humanities research: Wevers and Koolen (2020), for instance, present a study on word embeddings trained on a 500,000 digitized Dutch newspaper corpus for the purpose of studying the evolution of concepts.

3 Data and methods

Two newspaper collections were used for this experiment. A selection of titles from the British Library's *Heritage Made Digital* digitization project,[5] comprising 12 titles and around 2.3 billion tokens, and a collection specifically digitized for the Living with Machines project, comprising 107 titles and also around 2.3 billion tokens. Jointly, the collections span the period between 1801 and 1920. To prepare the corpora for training diachronic word embeddings, we first split them into time slices of 10 years each. We preprocessed the articles for each decade by removing word breaks resulting from OCR, newlines, and punctuation, by lowercasing the text and removing the stop words provided by the NLTK library for English.[6]

We trained Word2Vec (Mikolov et al., 2013) models as implemented in the Gensim library (Řehůřek and Sojka, 2010). To choose the optimal hyperparameters for training, we performed a grid search comparing the skip-gram and the continuous-bag-of-words architectures, as well as different number of epochs ({5,10}), vector dimensions ({200,300}), context windows ({3,5,10}) and minimum word counts ({1,5,10}). We evaluated the quality of the models resulting from all combinations of these parameters on one decade (all articles published between 1821 and 1830) calculating the cosine similarity between pairs of synonyms[7] in each model and choosing the model that returned the highest average score for all pairs. The final models were trained using the skip-gram architecture, 5 epochs, 200 dimensions, a context window of 3 and a minimum count of 1. Since the models for each decade are trained independently, the resulting word vectors in different decades are not aligned along the same coordinate axes. To allow for comparison between the representation of the same word across different decades, we aligned the semantic spaces on the basis of the Or-

[1] https://www.newseye.eu/
[2] https://translantis.wp.hum.uu.nl/
[3] https://impresso-project.ch/
[4] livingwithmachines.ac.uk

[5] https://www.bl.uk/projects/heritage-made-digital
[6] https://www.nltk.org/search.html?q=stopwords
[7] The list the synonyms considered is the following: *superfluous/unnecessary*, *display/exhibit*, *mimetic/imitative*, *disappear/vanish*, *alike/identical*.

The pairs were chosen so that at least one sense of one word is linked to a sense of the paired word via the linking between the Oxford English Dictionary and the Historical Thesaurus of English and the linked senses have quotations that include the range 1800-1920 or a portion of it.

thogonal Procrustes problem (Schönemann, 1966). Given $W^{(d)} \in \mathbb{R}^{n \times m}$, denoting the matrix of the vectors in decade d, the Orthogonal Procrustes problem consists in finding the orthogonal matrix $Q \in \mathbb{R}^{m \times m}$ that most closely maps the matrices $W^{(d)}$ and $W^{(d+1)}$. This is done by:

$$\min_{Q} ||W^{(d)}Q - W^{(d+1)}||_F, \quad \text{subject to } Q^TQ = I \tag{1}$$

where I is the $n \times m$ identity matrix and $||\dots||_F$ the Frobenius norm. The problem in (1) is solved via singular value decomposition: $U\Sigma V^T$, in this case $W^{(d)}(W^{(d+1)})^T$ (Tsakalidis et al., 2019, 2021). After all embedding spaces are aligned, we can use the cosine similarity between vectors across different decades to assess their semantic shift.

We compiled a list of words drawing from those indicated by Görlach (1999) as having undergone semantic change at some point during the 19th century.[8] For each word we calculated the cosine similarity between its vector in the semantic space for the most recent decade (the 1910s) and its vector in each of the previous decades. We followed Shoemark et al. (2019), who found that comparing the embeddings to the last time period leads to better results in semantic change detection. We then analysed the resulting scores in the following way. Any time point t with a cosine similarity significantly higher than the one in the time point $t-1$ was considered a potential changepoint in the meaning of a word. Significant changepoints were detected using the pruned exact linear time (PELT) algorithm (Killick et al., 2012), a penalized-cost method for detecting multiple changepoints in time-series data. We ran the algorithm with a jump parameter of 1 and comparing results with penalty set to 0.25 and 0.5.[9] We then extracted the nearest neighbours of each word for each decade to establish what type of semantic change might have occurred at each potential changepoint. We evaluated the accuracy of the models at detecting semantic change for a word against its entry in the Oxford English Dictionary (OED).[10] Using the OED API,[11] for each word we extracted the list of its senses, their definition and first record in writing, and selected all senses that had a first recorded year later than 1800 and earlier than 1920. To identify whether the detected potential changepoint for a word corresponded to one (or several) of its selected senses from the OED, we extracted the nearest neighbours of the word in each time period and compared those from the relevant decade(s) with the OED senses.

4 Qualitative analysis

word	changepoint
coach	1830
gear	1830
traffic	1830
train	1830
stamp	1840
fellow	1860
railway	1860
matches	1880

Table 1: Words with a changepoint detected by the PELT algorithm by setting the penalty to 0.5.

word	changepoint
wheel	1810, 1880
coach	1830
gear	1830
traffic	1830, 1860
train	1830
stamp	1840
fellow	1860
railway	1860
matches	1880

Table 2: Words with a changepoint detected by the PELT algorithm by setting the penalty to 0.25.

Table 1 contains the 8 words for which a semantic changepoint was detected by setting penalty to 0.5. As Table 2 shows, setting penalty to 0.25 resulted in detecting changepoints for one extra word.

[8] The complete list includes: *traffic, trade, train, coach, wheel, railway, matches, bulb, gear, stamp*. *Fellow* was also included as an example of semantically stable word made by Görlach (1999). For the purpose of this paper, words are considered only in their singular form for simplicity, even though considering both singulars and plurals may give a more complete picture. The only exception is the lemma *match*, which was considered only in its plural form, due to the intuitively more likely usage of this word in the plural (*matches*) in its new, phosphorous sense. Future studies may wish to consider both numbers for all the words and attempt reconciling, if needed, any different observations made on them.

[9] For this experiment we used the implementation of the PELT method by the `ruptures` library: https://pypi.org/project/ruptures/.

[10] https://www.oed.com

[11] https://languages.oup.com/research/oed-researcher-api/

We can immediately see that *fellow*, indicated by Görlach (1999, 131) as having a stable semantics in the 19th century, is included among the words in both tables. Two changepoints were also detected by the model trained with the lower penalty for *wheel*, another word cited by Görlach (1999, 131) as semantically stable. If we compare the trajectories of *wheel* and *train* (Figure 1), for example, it is not surprising to see that a changepoint detection model trained with stricter parameters may detect a change for the latter but not for the former, even though the plot suggests that a change in usage, albeit more gradual, occurred for *wheel* as well.

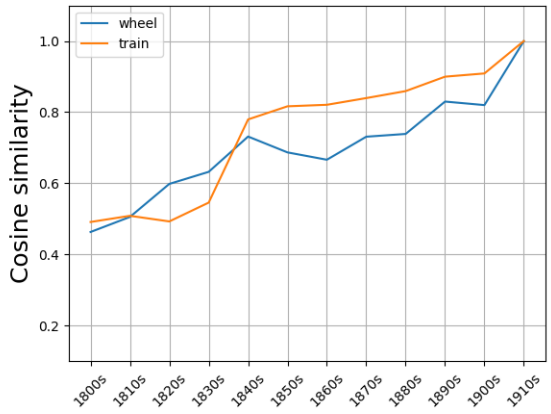

Figure 1: Time series for the cosine similarity between *wheel*, *train* in each decade and their respective vector in the time reference (the last decade, i.e. the 1910s).

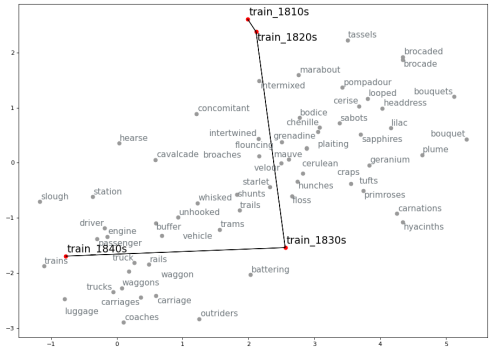

Figure 2: Semantic change trajectory of *train*.

4.1 *Train*

In Figure 2, we can see that *train* moved considerably in the semantic space between the 1810s and 1830s, to the extent that its 50-nearest neighbours in the 1810s and the 1840s have no words in common (see a selection of these in Table 3), with a decade in between, the 1820s, in which the words related to the older, more common sense ('an elongated back of a robe or skirt') are found together with those related to the newer one ('a series of connected railway carriages').[12] The semantics of this word appears to have changed steadily for at least two decades: our changepoint detection model was trained with a jump parameter of 1 (i.e. in our case, a change spanning at least one decade), so that a jump of 2 time units made it an even more likely candidate.

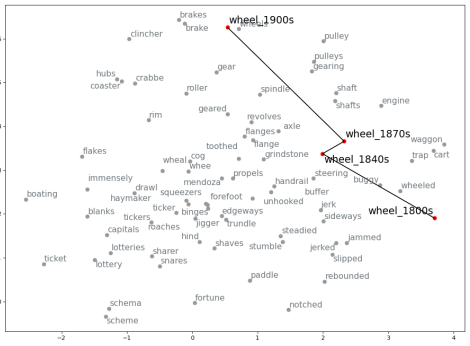

Figure 3: Semantic change trajectory of *wheel*.

4.2 *Wheel*

On the other hand, the semantic change of *wheel*, as suggested by our models, is less abrupt and may rather reflect an increased usage in specific senses related to technological innovations (or collocations describing these) throughout the century than the introduction of a new sense altogether, as was the case for *train*. If we compare the nearest neighbours of this word around the first changepoint (Table 3), we can see that words related to *wheel* in its figurative use referring to 'the course or sequence of events, procedure, the passage of time' prevail in the 1810s and 1820s, whereas words related to its sense 'various mechanical contrivances' are already the majority in the 1830s and 1840s. Terms related to the latter sense, however, are not exclusive of the period following the detected changepoint, as *carriage*, *cart*, *vehicle*, and *wagon* in the 1810s and 1820s all indicate. The OED lists the introduction of different specific usages of *wheel* in this

[12] Throughout the paper, the definition of the senses are quoted directly from the OED and reported in single quotation marks.

Word	Moving away from or adding new meanings to	Moving towards
coach	*saddle, harness, horses, post, telegraph* (1810s)	*wagon, driver, carriage, truck* (1830s)
fellow	*college, scholar, countrymen, bursar, tutor* (1850s)	*creatures, townsmen, countrymen, man, citizens* (1860s)
railway	*tunnel, turnpike, aqueducts, canals, navigation, drainage, waterworks* (1820s)	*railroad, junction, bridge, station, lines, tramway* (1830s); *beltway, companies, colliery, stakeholders, passengers, trains* (1860s)
traffic	*trafficking, slave, nefarious, kidnapping, illicit, contraband, smuggling, piracy* (1820s)	*railways, railroads, conveyance, transit, line* (1830s); *passengers, trains, coaches, milage* (1860s)
train	*chenille, intermixed, brocaded, lama, carnations* (1820s); *shunts, brocaded, mauve, hearse, carriages* (1830s)	*luggage, engine, carriages, waggons, trucks* (1840s)
wheel	*shafts, dray, stumble, draws, revolve, carriage, lottery, cart* (1810s); *drawn, dray, cart, prizes, shafts, capitals, vehicle, wagon* (1820s)	*axle, shaft, jerk, wheelers, flanges, jerked, axles, cart, paddle* (1830s); *axle, shaft, engine, buffer, flange, paddle, jammed* (1840s)

Table 3: Nearest neighbours of *coach*, *fellow*, *railway*, *traffic*, *train*, and *wheel* in the decades around the detected changepoints.

sense at different points in time since at least the 14th century, with *steering wheel* (1743) already in use in the nautical field and then extended to 'the steering-wheel of a motor vehicle'. A new usage of *wheel* recorded by the OED is that of *paddle wheel*, which appears among the nearest neighbours for the 1830s and 1840s (see Table 3), despite the OED reporting 1842 as its first written record. The clearest change between the 1820s and the 1830s is given by *train*-related words, such as *wagon* in the 1820s and *axle*, the closest neighbour of *wheel* in both the 1830s and 1840s.

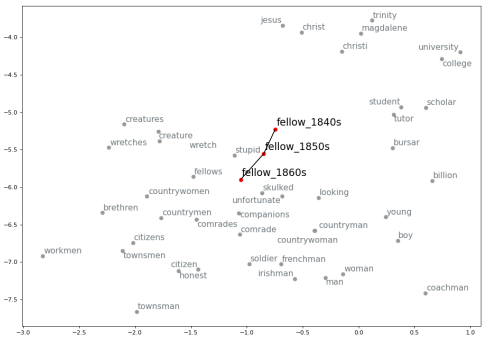

Figure 4: Semantic change trajectory of *fellow*.

4.3 *Fellow*

The case of *fellow* is also rather complex. By once again visualising the nearest neighbours for the detected changepoint and the preceding decades in a two-dimensional space, the neighbours are overall clearly divided between those related to *fellow* used in academic context (e.g. *tutor, scholar, college, bursar*, as names of specific colleges–*Magdalene, Trinity, Christi*), attested since the 15th century according to the OED, and those related to the sense of *fellow* broadly defined by the OED as 'a person who or thing which shares an attribute with another specified person or thing; a person or thing belonging to the same class or category as another' (e.g. *brethren, citizens, comrade, countrymen/countrywomen*), attested since the 13th century according to the OED. The OED however also records one new usage for the latter sense from 1844 ('a person's contemporary, esp. in a particular profession, art form, field of study, etc. chiefly in plural'), in addition to the similar, albeit more generic, pre-existing usage 'something that resembles another specified thing; a match; the like' for the same sense. Our models appear to reflect the new 1844 usage particularly in politically loaded words such as *citizens*, *brethren* and *comrade*, whose similarity with *fellow* may

be due to the political leanings of the newspapers in which this term appears the most. A new usage also recorded from 1816 by the OED is 'an animal or thing. Often affectionate, humorous, or ironic', which may be reflected in words such as *creatures*, the closest neighbour of *fellow* in the 1860s, as well as *unfortunate* and *wretch*.

4.4 *Railway* and *traffic*

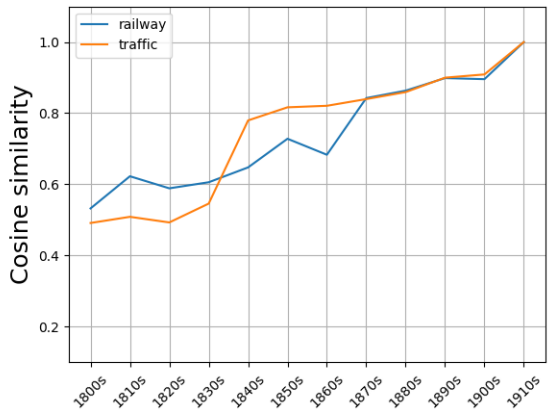

Figure 5: Time series for the cosine similarity between *railway* and *traffic* in each decade and their respective vector in the time reference (the last decade, i.e. the 1910s).

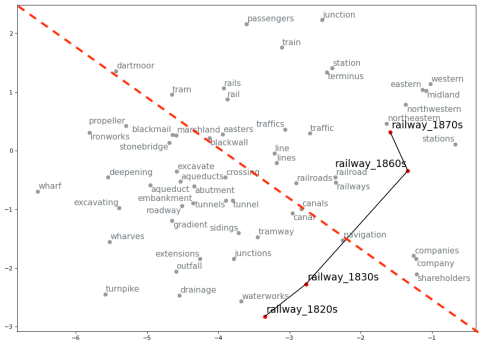

Figure 6: Semantic change trajectory of *railway*.

Other words from Tables 1 and 2 that pertain to the language of mechanization and that were mentioned by Görlach (1999) as examples of semantic change are *railway* and *traffic*. Two changepoints, the 1830s and the 1860s, were detected for *traffic* by the model trained with a lower penalty and both of these can be clearly seen in Figure 5. Only one changepoint, the 1860s, was instead detected for *railway*, as we can also gather from the steeper change in cosine similarity in the plot in Figure 5. However, it is quite evident that, besides the steep increase in cosine similarity between the 1860s and the 1870s, considerable change, though perhaps more gradual, occurred between the 1820s and the 1850s. This is in fact what we also observe if we compare the neighbours of *railway* before 1820 and after 1830 (Figure 6). A possible reason why no changepoint was detected pre-1860s is that its semantics up until the 1850s is not significantly dissimilar yet from the usage of the word in the previous two decades, when it may have been still widely used in the sense of 'a roadway laid with rails (originally of wood, later also of iron or steel) along which the wheels of wagons or trucks may run, in order to facilitate the transport of heavy loads'. A neater departure from the latter is observed by the 1860s, when it was probably already used predominantly in the sense of 'a line or track typically consisting of a pair of iron or steel rails, along which carriages, wagons, or trucks conveying passengers or goods are moved by a locomotive engine or other powered unit', first attested, according to the OED, in the 1820s. Between the 1830s and the 1860s, the key change in the meaning of *railway*, which can be inferred from the semantic space in Figure 6, is two-fold. First, there is a definite departure from railways as only a means for the transport of goods to railways as a means of transportation for passengers. This is evident from the distance of *railway* in the 1860s from the words in Figure 6 concerning precisely this semantic field, such as *canals, tunnel, navigation, waterworks, excavating, wharf, embankment, roadway, turnpike* or *aqueducts*, and the greater proximity to words such as *train, station, passengers* and *tram*. The proximity to these latter words is particularly clear by focussing on the axis highlighted with a red dashed line in Figure 6, across which the semantic change seems to have occurred. Second, we observe a shift towards the usage of *railway* in the meaning of 'a network or organization of such lines [as defined by the new sense defined of railway above]; a company which owns, manages, or operates such a line or network; this form of transportation'. This is clear from neighbours such as *company* and *shareholders*, and modifiers that were likely to identify clearly defined regional railway networks, such as *northwestern*, *midland*, and *western*.

Both changepoints for the word *traffic* are

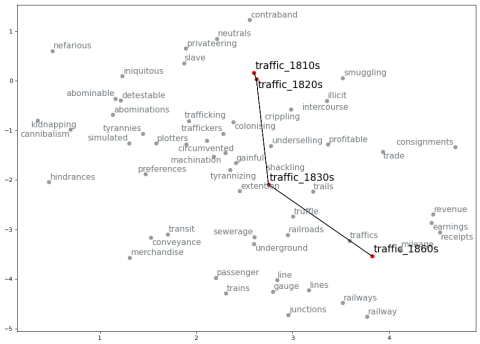

Figure 7: Semantic change trajectory of *traffic*.

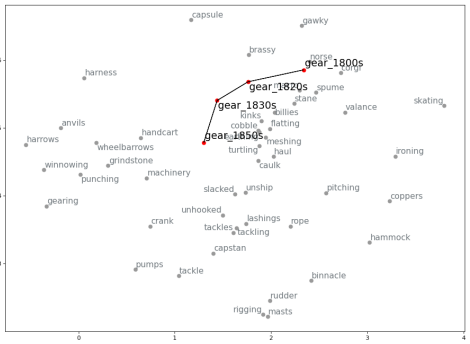

Figure 8: Semantic change trajectory of *gear*.

supported by our neighbour analysis. Between the 1820s and the 1830s the main meaning of *traffic* drifted away from the sense defined by the OED as 'the activity or business of acquiring, transporting, and selling something which, for legal or moral reasons, should not be treated as a mere commodity; trade of an illegal, immoral, or otherwise objectionable nature', exemplified by 1810s-1820s nearest-neighbours such as *slave, contraband, detestable, infamous, inhuman, abominable, execrable, disgraceful, trafficking*[13], *piracy, illicit,* and so on. Its main usage by the 1830s, as suggested by its neighbours, is in the sense of 'passage of vehicles, vessels, etc., to and fro along a route', and by the 1860s several neighbours are related to its usage (first recorded, according to the OED, in the 1830s) as 'the quantity of goods, or number of passengers, carried by a transportation service over a particular period; the business or revenue generated from this', as exemplified by words such as *passengers, coaches, railways, trains* and *milage*.

4.5 *Gear*

The trajectory of *gear* (Figure 8) is exemplary of a general trend towards specific senses related to new mechanical advances throughout the 19th century, reflecting the several new usages related to 'machinery' recorded by the OED as first being attested at different points between the 1810s and the 1870s.

[13]This word specifically is indicated by the OED as an example of traffic in this sense.

4.6 *Matches* and *stamp*

The words *matches* and *stamp*, for both of which a potential changepoint was detected by the model trained with a lower penalty, were mentioned by Görlach (1999, 128) when noting that Soule (1871) in his *A Dictionary of English Synonymes* failed to include the 'phosphorous sense of match [and] the philatelic sense of stamp', which Görlach explains as possibly due to the fact the new senses had not become dominant in the 19th century yet.

Our results for *stamp* (Figure 9), however, suggests that by the 1860s the philatelic sense (first attested according to the OED in 1837) was already prominent, as we can see from words such as *envelope, postage*, and *penny* (possibly referring to the price of a stamp), unlike the nearest neighbours of the word in the 1840s, such as *affixing, engrave, government*, or *grave*, which are related to the main older sense of *stamp* as 'the mark, impression, or imprint made with an engraved block or die'.

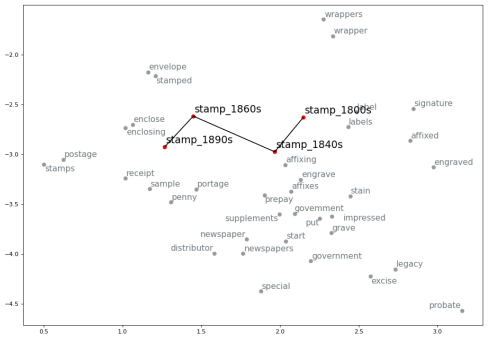

Figure 9: Semantic change trajectory of *stamp*.

Unlike *stamp*, the results of our changepoint detection method for *matches* are likely to be mislead-

ing and could be heavily biassed by a particular event (possibly sports-related) being heavily covered by the news between the 1880s and 1890s. In Figure 10 we can see that, although the new 'phosphorous sense' of the word is among the nearest neighbours in the plot (e.g. *phosphorus* and *ignite*), their cosine similarity with *match* is likely not as high between the 1860s and the 1880s (the period within which a potential changepoint was detected) as that with words related to the pre-existing sense 'a contest or competitive trial of skill in a particular sport, game, or other activity'.

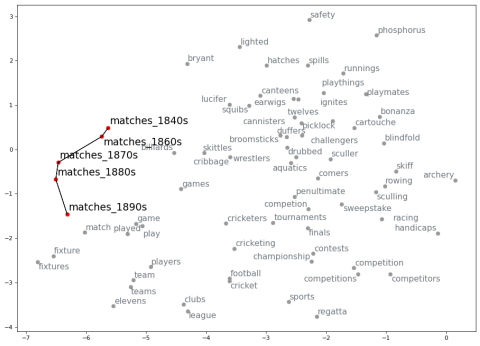

Figure 10: Semantic change trajectory of *matches*.

4.7 *Coach*

Coach is discussed by Görlach (1999, 128) as having undergone semantic extension from its meaning as a 'large horse-drawn carriage', attested since the 16th century, to the sense, recorded in the OED, 'a railway carriage', an extension which is also visible from the semantic space of this word and its neighbours from our diachronic models (Figure 11). This is an especially encouraging result, since our models captured this semantic extension as early as the decade recorded by the OED as the first written attestation, while also showing that its usage in the first half of the 19th century was not exclusively American English as defined in the OED and reported by Görlach (1999, 128).

A possible explanation as to why for words like *bulb* no definite changepoint was detected is that the semantic change trajectory of such words may be much more complex than a mere addition of a sense and a significant spread in use of the latter around a specific decade. Specifically in the case of *bulb*, according to the OED, at least three main senses were already in use at the beginning of the 19th century from different semantic fields

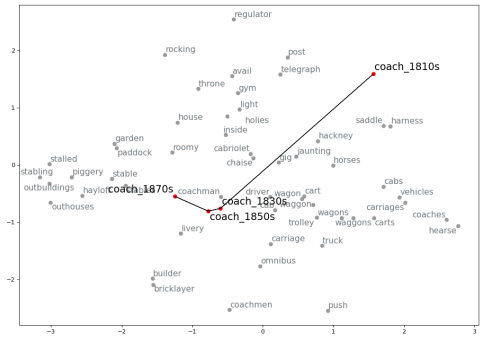

Figure 11: Semantic change trajectory of *coach*.

(anatomy, botany, and, broadly, electricity). New specific uses of the word are then attested from the mid-19th century, but these are classified by the OED as specialisations of two of the previously existing senses, sometimes specifically when the words are found within certain collocations (as in *electric light bulb*, first recorded in 1856 according to the OED). Görlach (1999, 134) mentions *bulb*, together with *gear* and *stamp*, as examples of words that underwent 'conspicuous semantic changes caused by technological progress', comparing the expansion of meaning of these words to that of *circuit* and *current* towards their electricity-related sense in the previous century. It is useful to note that overall trajectory of *bulb*, *gear* and *stamp* appears to be quite similar (Figure 12).

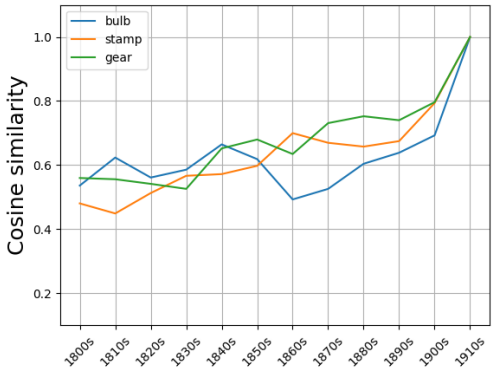

Figure 12: Time series for the cosine similarity between *bulb*, *gear*, *stamp* in each decade and their respective vector in the time reference (the last decade, i.e. the 1910s).

Although the general trajectory is slightly upward (i.e. there is likely an overall change in meaning) for all three words, *stamp* and *gear* show a

more gradual but somewhat steadier change in cosine similarity with the vector of the reference time period (1910s), starting from a cosine similarity below 0.6 and reaching 0.8, a very high score, towards the beginning of the 20th century. *Bulb*, on the other hand, has a less regular trajectory and hardly reaches a cosine similarity with its 1900s representation of 0.7.

5 Quality control

Since large digitized newspaper collections are frequently not created with a specific criterion in mind, but rather following specific policies of the digitizing institution, we needed to be particularly wary that the likely biassed content of our data (cf. Beelen et al., 2022) would not significantly interfere with our research questions. The quality of our models and validity of our method were checked in several ways.

First, to make sure that potential detected changepoints were not simply the result of a biassed dataset, we ran our changepoint detection method individually on all the words in the list of synonym pairs which were also used to optimize the embedding hyperparameters, since they were indicated by the OED as semantically stable throughout our period of interest. With a jump parameter of 1 and a penalty of 0.5 (the safer, stricter option), no changepoint was detected for any of the words, with the exception of *identical*, suggesting an overall good reliability for our models.

Second, throughout the analysis we used two external sources to validate our results. *A history of the English Language in the Nineteenth Century* by Görlach (1999), specifically its chapter on lexical change, was used to draw examples from the language of mechanization that the scholar indicated as having undergone some type of semantic change. We also included words which he mentioned as seeming semantically stable throughout the century (namely *fellow* and *wheel*, the former not in the lexical field of mechanization) as a further form of comparison with non-digital scholarship on the subject. Finally, throughout the analysis we employed the OED as a benchmark to check whether a changepoint coincided with a newly recorded senses, as well as to identify definition of new senses and usages, especially in highly polysemous or ambiguous contexts.

6 Conclusions

In this paper we presented a first attempt at a large-scale computational study of semantic change of terms related to the lexical field of mechanisation in 19th-century English. Our main goal was to find out whether vector space models trained on very large (4.6B tokens) digitized, hence noisy, historical newspapers were able to stand the test of expert knowledge on the topic. We showed that using changepoint detection methods on the diachronic word embeddings that we trained gave results most often matching the observations made by traditional scholarship. Through a combination of changepoint detection and neighbour analysis it was possible to provide explanations for mismatches between previous literature and our findings, in some cases noticing that our results were able to capture features of semantic change not identified by the expert sources (see, for example, the analysis of *coach* above).

Our analysis provides the bases for new data-driven investigations on the lexical field of mechanization that do not rely so closely on external knowledge bases as in our study.

Acknowledgments

Work for this paper was carried out as part of *Living with Machines*. This project, funded by the UK Research and Innovation (UKRI) Strategic Priority Fund, is a multidisciplinary collaboration delivered by the Arts and Humanities Research Council (AHRC grant AH/S01179X/1), with The Alan Turing Institute, the British Library and King's College London, East Anglia, Exeter, and Queen Mary University of London. Newspaper digitisation by the British Library, undertaken as part of the Living with Machines project, was funded by the AHRC.

References

Kaspar Beelen, Jon Lawrence, Daniel C. S. Wilson, and David Beavan. 2022. Bias and representativeness in digitized newspaper collections: Introducing the environmental scan. *Digital Scholarship in the Humanities*, pages 1–22.

Alexander Bergs and Laurel J. Brinton, editors. 2012. *English Historical Linguistics*. Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 34.1. De Gruyter Mouton, Berlin, Boston.

Ingegerd Bäcklund. 2006. *Modifiers describing women*

and men in nineteenth-century English, pages 17–55. Cambridge University Press, Cambridge.

Manfred Görlach. 1999. *English in Nineteenth-Century England: An Introduction*. Cambridge University Press, Cambridge.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Christian Kay and Kathryn Allan. 2015. *English Historical Semantics*. Edinburgh University Press, Edinburgh.

Roberta Killick, Paul Fearnhead, and Idris A. Eckley. 2012. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

M. Kytö, M. Rydén, and E. Smitterberg, editors. 2006. *Nineteenth-century English: Stability and change*. Cambridge University Press, Cambridge.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Lynda Mugglestone. 2008. *The Oxford History of English*. Oxford University Press, Oxford.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Peter H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31:1–10.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

Richard Soule. 1871. *A Dictionary of English Synonymes and Synonymous Or Parallel Expressions: Designed as a Practical Guide to Aptness and Variety of Phraseology*.

Adam Tsakalidis, Piero Basile, Marya Bazzi, Mihai Cucuringu, and Barbara McGillivray. 2021. DUKweb, diachronic word representations from the UK Web Archive corpus. *Scientific Data*, 8(269).

Adam Tsakalidis, Marya Bazzi, Mihai Cucuringu, Pierpaolo Basile, and Barbara McGillivray. 2019. Mining the UK web archive for semantic change detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1212–1221, Varna, Bulgaria. INCOMA Ltd.

Daniel van Strien, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the Impact of OCR Quality on Downstream NLP Tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1: ARTIDIGH*, pages 484–496. INSTICC, SciTePress.

Melvin Wevers and Marijn Koolen. 2020. Digital begriffsgeschichte: Tracing semantic change using word embeddings. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(4):226–243.

A Scripts and models

All the Python scripts used to train the diachronic word embeddings, as well as several Jupyter notebooks to replicate the methodology employed in this paper, can be found at https://github.com/Living-with-machines/DiachronicEmb-BigHistData.

The vectors used for the analysis in this paper can be found in the following repository in Zenodo: https://doi.org/10.5281/zenodo.7181681.

B Visualization method

Visualization of the semantic trajectories is carried out in the following steps:

1. Define three or four decades around which a semantic shift appears to have taken place for a word *w*. This is established through a combination of automatic changepoint detection and close reading of the neighbours of *w*. The selected decades should be adjusted across different runs to achieve the clearest visual rendition of a semantic shift (if any).

2. Extract the 20-nearest neighbours of *w* for the selected decades and remove any duplicate

(i.e. neighbours of *w* appearing in more than one decade).

4. From the extracted neighbours, remove words that are clear misspellings (likely due to OCR errors).

5. From the model for the most recent decade (among the selected decades) extract the vector of each word in the list of neighbours. Discard words that are not in the vocabulary of the model.

6. Add the vectors for *w* from each of the selected decades to resulting list of vector and convert this list to a `numpy` array.

7. Reduce dimensionality using T-distributed Stochastic Neighbor Embedding (t-SNE)[14]

8. Visualize the resulting two-dimensional embedding space in a scatter plot, highlighting the label for *w* in the selected decades. For details on the latter, see the code repository.

[14] To do this, we used the implementation of t-SNE by the `sklearn` library, setting the number of dimensions to `2`, the maximum number of iterations to `1000`, and the initialization method to `random`.