WORD EMBEDDINGS

FOR THE ANALYSIS OF DIATOPIC & DIACHRONIC SEMANTIC VARIATION FROM HISTORICAL NEWSPAPERS

AIUCD 2023, SIENA, ITALY
NILO PEDRAZZINI
BARBARA MCGILLIVRAY

BACKGROUND

- ✔ The British *Industrial Revolution* was a time of profound changes and **mechanization** was its primary driver.
- ✔ New machines and new technology heavily affected the English lexicon: new words are coined and existing words get new meanings.

CHALLENGE

- Given large enough datasets, can we **automatically detect changes** in the meaning of words associated with this process?
- Can we assess the extent to which **diatopic variation** is an important variable in these changes?

DATA & METHODS

- ⇛ 19th-century (1800-1920) **British newspaper** corpus (4.5B tokens), including the Heritage Made Digital (HMD) and Living with Machines (LwM) collections. Metadata on place of publication used to divide corpus into geographical regions (North and South England, Midlands, Scotland and Wales).
- ⇛ Aligned diachronic word embeddings (**Word2Vec**) trained on this corpus (Hamilton et al. 2016).
- ⇛ Change point detection (**PELT** algorithm; Killick et al. 2012).
- ⇛ Qualitative analysis and validation against **traditional scholarship** (Görlach 1993): testing methods on *car*(*s*), *bike*(*s*), *trolley*(*s*), *bus*(*es*), *tram*(*s*), *machine*(*s*), *traffic*, *trade*(*s*), *train*(*s*), *coach*(*es*), *wheel*(*s*), *railway*(*s*), *matche*(*s*), *bulb*(*s*), *gear*(*s*), *stamp*(*s*)

TRAINING & ALIGNMENT

- ⇄ Grid search performed to find optimal hyperparameters.
- ⇄ One model per decade was trained with Word2Vec (Gensim), SkipGram, 5 epochs, 200 dimensions, window of 5 and minimum count of 1. The process was repeated for each geographical region.
- ⇄ All semantic spaces within a region were aligned (with Orthogonal Procrustes) to the most recent time slice.

CASE STUDY

- ⇛ As a case application of the diachronic and diatopic word embeddings, we consider potential differences in the semantic change of words related to the lexicon of mechanization between the North and South of England, a historically major split in socio-political terms in 19th-century Britain.
- ⇛ The PELT algorithm (0.25 penalty, 1 jump) detected changepoints for different words in the two regions, while for other words a changepoint was detected for both, but in different decades.

Word	North England	South England
bulbs	1860s	-
cars	1860s	-
machines	-	1860s
match	1860s	1840s
matches	1860s	-
stamp	1860s	1860s
stamps	-	1840s
stock	-	1860s
trade	1860s	-
trolley	*not in vocabulary*	1850s

- ⇛ Changepoint detection suggests that the semantic change of words related to the lexicon of mechanization did not occur at the same pace across British regions.
- ⇛ To evaluate the type of semantic change detected by the potential changepoints, we can analyse the *k*-nearest neighbours of the words undergoing potential semantic change. Upon close inspection, some words may have either undergone change before our period of interest or might have had a less sudden change in usage. The word ***cars***, for example, appears to have had a sudden change in usage in North England from its older existing meaning of *a wheeled, usually horse-drawn conveyance* [OED] to that associated with railway carriages or wagons (Figure below, left), unlike the South, where the newer meaning is also attested but does induce a changepoint detectable by the algorithm.
- ⇛ The difference in changepoints between North and South for words like ***stock*** and ***trade***, as well the differences between singular and plural usages of the words can provide interesting insights to historians researching regional differences during the time following the Industrial Revolution.

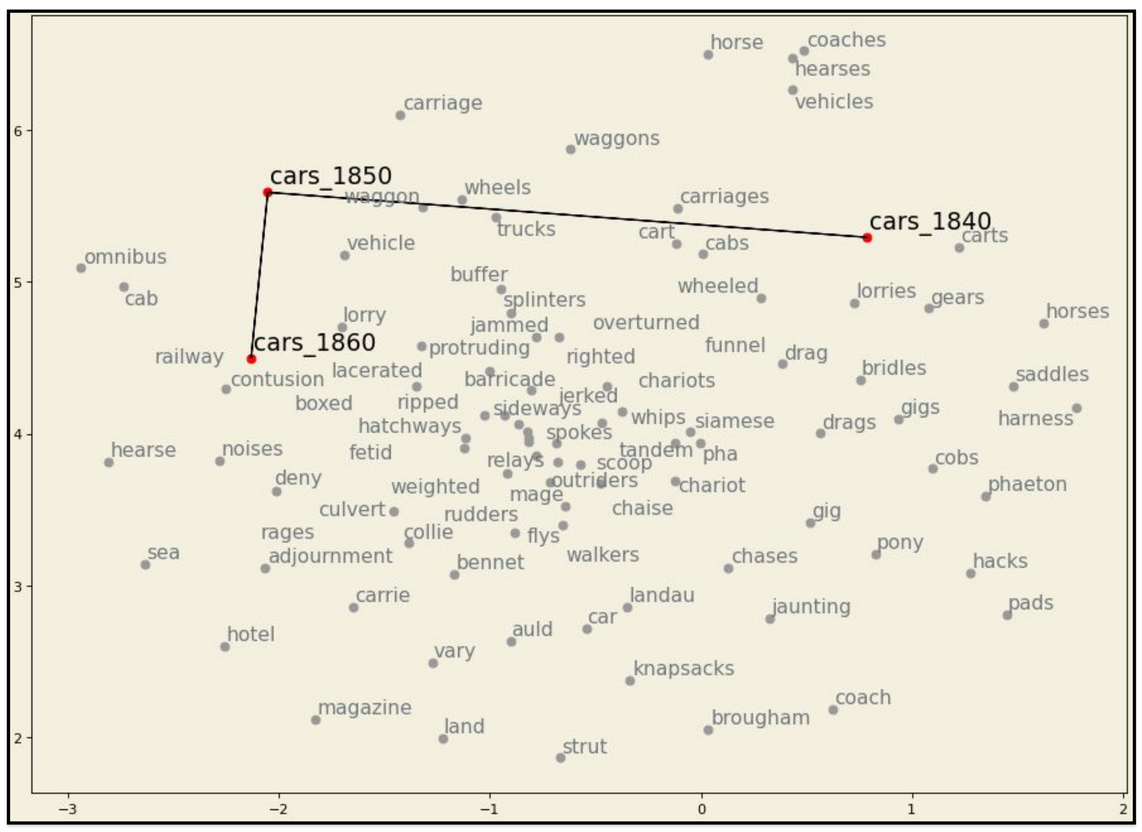

Figure 1. Semantic change trajectory for *cars* in North England, based on its nearest neighbours in three decades.

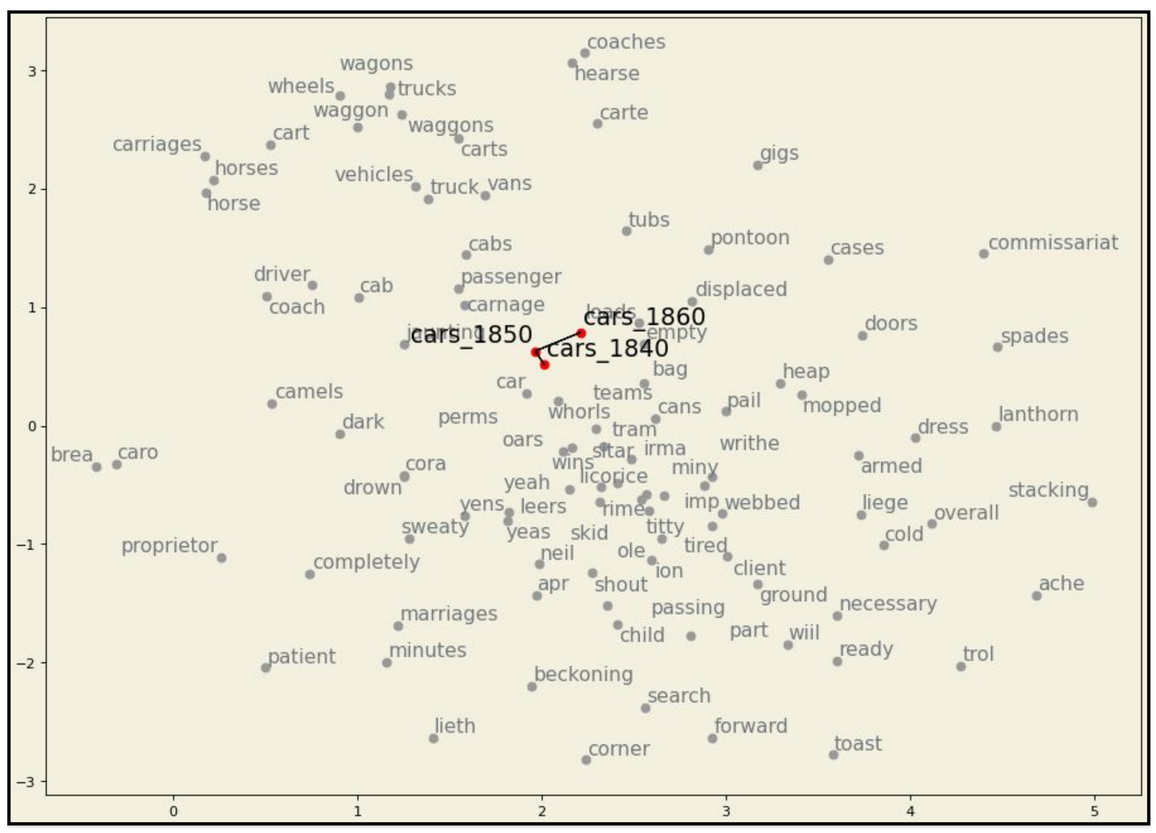

Figure 2. Semantic change trajectory for *cars* in North England, based on its nearest neighbours in three decades.

RESOURCES

- ★ **Code**
  https://github.com/Living-with-machines/DiachronicEmb-BigHistData
- ★ **Models**
  https://doi.org/10.5281/zenodo.7892460

REFERENCES

- ★ **Görlach 1993:** Görlach, M. 1999. *English in Nineteenth-Century England: An Introduction*. Cambridge: Cambridge University Press.
- ★ **Hamilton et al. 2016:** Hamilton, W.L., J. Leskovec & D. Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1489–1501.
- ★ **Killick et al. 2012:** Killick, R., P. Fearnhead & I.A. Eckley. 2012. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500), 1590–1598.
- ★ **Pedrazzini & McGillivray 2022:** Pedrazzini, N., and B. McGillivray. 2022. Machines in the media: semantic change in the lexicon of mechanization in 19th-century British newspapers. *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, 85–95.