Embeddings Shifts as Proxies for Different Word Use in Italian Newspapers

Michele Cafagna[1,3], **Lorenzo De Mattei**[1,2,3] **and Malvina Nissim**[3]

[1]Department of Computer Science, University of Pisa, Italy
[2]ItaliaNLP Lab, ILC-CNR, Pisa, Italy
[3]University of Groningen, The Netherlands
{m.cafagna,m.nissim}@rug.nl, {lorenzo.demattei}@di.unipi.it

Abstract

We study how words are used differently in two Italian newspapers at opposite ends of the political spectrum by training embeddings on one newspaper's corpus, updating the weights on the second one, and observing vector shifts. We run two types of analysis, one top-down, based on a preselection of frequent words in both newspapers, and one bottom-up, on the basis of a combination of the observed shifts and relative and absolute frequency. The analysis is specific to this data, but the method can serve as a blueprint for similar studies.

1 Introduction and Background

Different newspapers, especially if positioned at opposite ends of the political spectrum, can render the same event in different ways. In Example (1), both headlines are about the leader of the Italian political movement "Cinque Stelle" splitting up with his girlfriend, but the Italian left-oriented newspaper *la Repubblica*[1] (`rep` in the examples) and right-oriented *Il Giornale*[2] (`gio` in the examples) describe the news quite differently. The news in Example (2), which is about a baby-sitter killing a child in Moscow, is also reported by the two newspapers mentioning and stressing different aspects of the same event.

(1) `rep` La ex di Di Maio: "E' stato un amore intenso ma non abbiamo retto allo stress della politica"
[*en: The ex of Di Maio: "It's been an intense love relationship, but we haven't survived the stress of politics"*]

`gio` Luigino single, è finita la Melodia
[*en: Luigino single, the Melody is over*]

(2) `rep` Mosca, "la baby sitter omicida non ha agito da sola"
[*en: Moscow, "the killer baby-sitter has not acted alone"*]

`gio` Mosca, la donna killer: "Ho decapitato la bimba perché me l'ha ordinato Allah"
[*en: Moscow, the killer woman: "I have beheaded the child because Allah has ordered me to do it"*]

Often though, the same words are used, but with distinct nuances, or in combination with other, different words, as in Examples (3)–(4):

(3) `rep` Usa: agente uccide un nero disarmato e immobilizzato
[*en: Usa: policeman kills an unarmed and immobilised black guy*]

`gio` Oklahoma, poliziotto uccide un nero disarmato: "Ho sbagliato pistola"
[*en: Oklahoma: policeman kills an unarmed black guy: "I used the wrong gun"*]

(4) `rep` Corte Sudan annulla condanna, Meriam torna libera
[*en: Sudan Court cancels the sentence, Meriam is free again*]

`gio` Sudan, Meriam è libera: non sarà impiccata perché cristiana
[*en: Sudan: Meriam is free: she won't be hanged because Christian*]

In this work we discuss a method to study how the same words are used differently in two sources, exploiting vector shifts in embedding spaces.

The two embeddings models built on data coming from *la Repubblica* and *Il Giornale* might contain interesting differences, but since they are separate spaces they are not directly comparable. Previous work has encountered this issue from a diachronic perspective: when studying meaning shift in time, embeddings built on data from different periods would encode different usages, but they need to be comparable. Instead of constructing separate spaces and then aligning them

[1] https://www.repubblica.it

[2] http://www.ilgiornale.it

(Hamilton et al., 2016b), we adopt the method used by Kim et al. (2014) and subsequently by Del Tredici et al. (2016) for Italian, whereby embeddings are first trained on a corpus, and then updated with a new one; observing the shifts certain words undergo through the update is a rather successful method to proxy meaning change.

Rather than across time, we update embeddings across sources which are identical in genre (newspapers) but different in political positioning. Specifically, we train embeddings on articles coming from the newspaper *La Repubblica* (leaning left) and update them using articles coming from the newspaper *Il Giornale* (leaning right). We take the observed shift of a given word (or the shift in distance between two words) as a proxy for a difference in usage of that term, running two types of analysis. One is top-down, and focuses on a set of specific words which are frequent in both corpora. The other one is bottom-up, focusing on words that result potentially interesting on the basis of measures that combine the observed shift with both relative and absolute frequency. As a byproduct, we also learn something about the interaction of shifts and frequency.

2 Data

We scraped articles from the online sites of the Italian newspapers *la Repubblica*, and *Il Giornale*. We concatenated each article to its headline, and obtained a total of 276,120 documents (202,419 for *Il Giornale* and 73,701 for *la Repubblica*).

For training the two word embeddings, though, we only used a selection of the data. Since we are interested in studying how the usage of the same words changes across the two newspapers, we wanted to maximise the chance of articles from the two newspapers being on the same topic. Thus, we implemented an automatic alignment, and retained only the aligned news for each of the two corpora. All embeddings are trained on such aligned news.

2.1 Alignment

We align the two datasets using the whole body of the articles. We compute the tf-idf vectors for all the articles of both newspapers and create subsets of relevant news filtering by date, i.e. considering only news that were published in the range of three days before and after of one another. Once this subset is extracted, we compute cosine similarities for all news in one corpus and in the other corpus using the tf-idf vectors, we rank them and then filter out alignments whose cosine similarity is under a certain threshold. The threshold should be chosen taking into consideration a trade-off between keeping a sufficient number of documents and quality of alignment. In this case, we are relatively happy with a good but not too strict alignment, and after a few tests and manual checks, we found that threshold of 0.185 works well in practice for these datasets, yielding a good balance between correct alignments and news recall. Table 1 shows the size of the aligned corpus in terms of number of documents and tokens.

newspaper	**#documents**	**#tokens**
la Repubblica	31,209	23,038,718
Il Giornale	38,984	18,584,121

Table 1: Size of the aligned corpus.

2.2 Shared lexicon

If we look at the most frequent content words in the datasets (Figure 1), we see that they are indeed very similar, most likely due to the datasets being aligned based on lexical overlap.

This selection of frequent words already constitutes a set of interesting tokens to study for their potential usage shift across the two newspapers. In addition, through the updating procedure that we describe in the next section, we will be able to identify which words appear to undergo the heaviest shifts from the original to the updated space, possibly indicating a substantial difference of use across the two newspapers.

2.3 Distinguishability

Seeing that frequent words are shared across the two datasets, we want to ensure that the two datasets are still different enough to make the embeddings update meaningful.

We therefore run a simple classification experiment to assess how distinguishable the two sources are based on lexical features. Using the scikit-learn implementation with default parameters (Pedregosa et al., 2011), we trained a binary linear SVM to predict whether a given document comes from *la Repubblica* or *Il Giornale*. We used ten-fold cross-validation over the aligned dataset with only word n-grams 1-2 as features and obtained an overall accuracy of 0.796, and 0.794 and 0.797 average precision and recall, respectively.

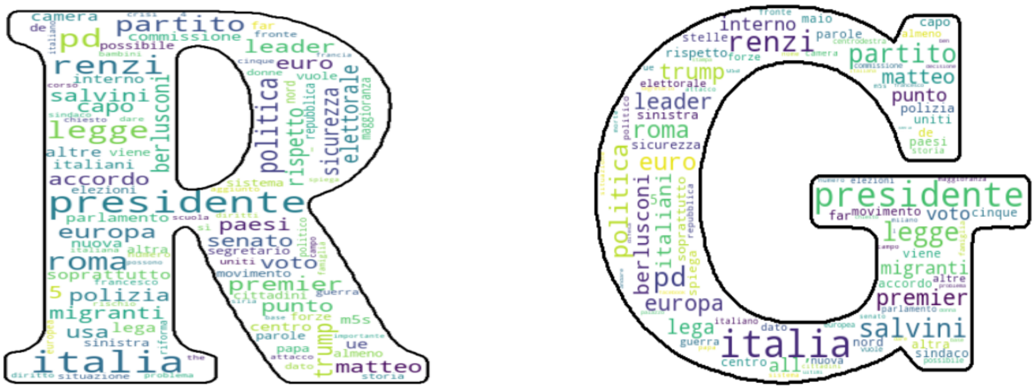

Figure 1: Left: top 100 most frequent words in *la Repubblica*. Right: top 100 in *Il Giornale*.The words are scaled proportionally to their frequency in the respective datasets.

This is indicative that the two newspapers can be distinguished even when writing about the same topics. Looking at predictive features we can indeed see some words that might be characterising each of the newspapers due to their higher tf-idf weight, thus maintaining distinctive context even in similar topics and with frequent shared words.

3 Embeddings and Measures

We train embeddings on one source, and update the weights training on the other source. Specifically, using the gensim library (Řehůřek and Sojka, 2010), first we train a word2vec model (Mikolov et al., 2013) to learn 128 sized vectors on *la Repubblica* corpus (using the skip-gram model, window size of 5, high-frequency word downsample rate of 1e-4, learning rate of 0.05 and minimum word frequency 3, for 15 iterations). We call these word embeddings $spaceR$. Next, we update $spaceR$ on the documents of *Il Giornale* with identical settings but for 5 iterations rather than 15. The resulting space, $spaceRG$, has a total vocabulary size of 53,684 words. We decided to go this direction (rather than train on Il Giornale first and update on La Repubblica later because the La Repubblica corpus is larger in terms of tokens, thus ensuring a more stable space to start from.

3.1 Quantifying the shift

This procedure makes it possible to observe the shift of any given word, both quantitatively as well as qualitatively. This is more powerful than building two separate spaces and just check the nearest neighbours of a selection of words. In the same way that the distance between two words is approximated by the cosine distance of their vectors (Turney and Pantel, 2010), we calculate the distance between a word in $spaceR$ and the same word in $spaceRG$, by taking the norm of the difference between the vectors. This value for word w is referred to as $shift_w$. The higher $shift_w$, the larger the difference in usage of w across the two spaces. We observe an average shift of 1.98, with the highest value at 6.65.

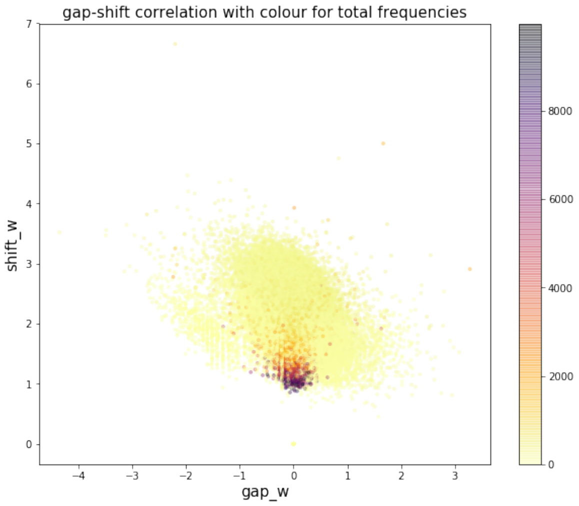

Figure 2: Gap-Shift scatter plot of the words in the two newspapers. Darker colour indicates a higher cumulative frequency; a negative gap means higher relative frequency in *Il Giornale*.

3.2 Frequency impact

By looking at raw shifts, selecting high ones, we could see some potentially interesting words.

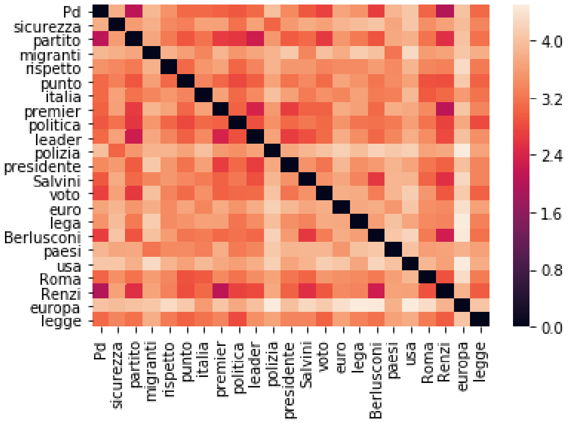

Figure 3: Distance matrix between a small set of high frequency words on *la Repubblica*. The lighter the color the larger the distance.

However, frequency plays an important role, too (Schnabel et al., 2015). To account for this, we explore the impact of both absolute and relative frequency for each word w. We take the overall frequency of a word summing the individual occurrences of w in the two corpora ($total_w$). We also take the difference between the relative frequency of a word in the two corpora, as this might be influencing the shift. We refer to this difference as gap_w, and calculate it as in Equation 1.

$$(1) \qquad gap_w = log(\frac{freq_w^r}{|r|}) - log(\frac{freq_w^g}{|g|})$$

A negative gap_w indicates that the word is relatively more frequent in *Il Giornale* than in *la Repubblica*, while a positive value indicates the opposite. Words whose relative frequency is similar in both corpora exhibit values around 0.

We observe a tiny but significant negative correlation between $total_w$ and $shift_w$ (-0.093, $p < 0.0001$), indicating that the more frequent a word, the less it is likely to shift. In Figure 2 we see all the dark dots (most frequent words) concentrated at the bottom of the scatter plot (lower shifts).

However, when we consider gap_w and $shift_w$, we see a more substantial negative correlation (-0.306, $p < 0.0001$), suggesting that the gap has an influence on the shift: the more negative the gap, the higher the shift. In other words, the shift is larger if a word is relatively more frequent in the corpus used to update the embeddings.

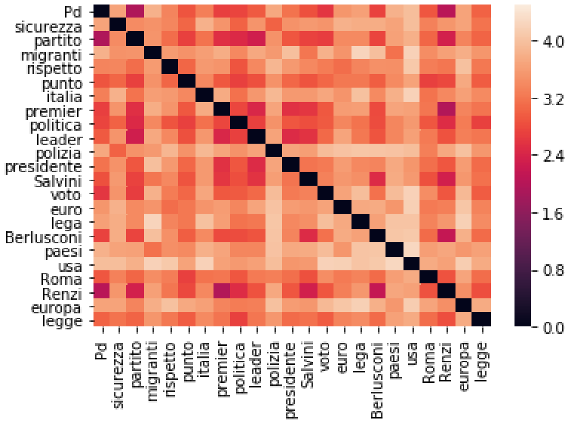

Figure 4: Distance matrix between a small set of high frequency words after updating with *Il Giornale*. The lighter the color the larger the distance.

4 Analysis

We use the information that derives from having the original $spaceR$ and the updated $spaceRG$ to carry out two types of analysis. The first one is top-down, with a pre-selection of words to study, while the second one is bottom-up, based on measures combining the shift and frequency.

4.1 Top-down

As a first analysis, we look into the most frequent words in both newspapers and study how their relationships change when we move from $spaceR$ to $spaceRG$. The words we analyse are the union of those reported in Figure 1. Note that in this analysis we look at pairs of words at once, rather than at the shift of a single word from one space to the next. We build three matrices to visualise the distance between these words.

The first matrix (Figure 3) only considers $SpaceR$, and serves to show how close/distant the words are from one another in *la Repubblica*. For example, we see that “partito” and “Pd”, or “premier” and “Renzi” are close (dark-painted), while “polizia” and “europa” are lighter, thus more distant (probably used in different contexts).

In Figure 4 we show a replica of the first matrix, but now on $SpaceRG$; this matrix now let’s us see how the distance between pairs of words has changed after updating the weights. Some vectors are farther than before and this is visible by the ligther color of the figure, like “usa” and “lega” or “italia” and “usa”, while some words are closer like “Berlusconi” and “europa” or “europa” and “politica” which feature darker colour. Specific

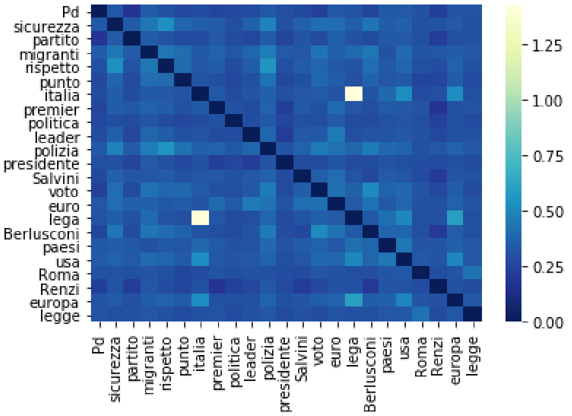

Figure 5: Difference matrix between embeddings from *spaceR* and *spaceRG* normalised with the logarithm of the absolute frequency difference in *spaceRG*. The lighter the colour, the larger the distance between pairs of words.

analysis of the co-occurrences of such words could yield interesting observations on their use in the two newspapers.

In order to better observe the actual difference, the third matrix shows the shift from $spaceR$ to $spaceRG$, normalised by the logarithm of the absolute difference between the $total_{w1}$ and $total_{w2}$ (Figure 5).[3] Lighter word-pairs shifted more, thus suggesting different contexts and usage, for example “italia” and “lega”. Darker pairs, on the other hand, such as “Pd”-“Partito” are also interesting for deeper analysis, since their joint usage is likely to be quite similar in both newspapers.

4.2 Bottom-up

Differently from what we did in the top-down analysis, here we do not look at how the relationship between pairs of pre-selected words changes, rather at how a single word’s usage varies across the two spaces. These words arise from the interaction of gap and $shift$, which yields various scenarios. Words with a large negative gap (relative frequency higher in *Il Giornale*) are likely to shift more, but it’s probably more of an effect due to increased frequency than a genuine shift. Words that have a high gap (occurring relatively less in *Il Giornale*) are likely to shift less, most likely since adding a few contexts might not cause much shift.

The most interesting cases are words whose relative frequency does not change in the two datasets, but have a high shift. Zooming in on the words that have small gaps ($-0.1 < gap_w < 0.1$), will provide us with a set of potentially interesting words, especially if they have a shift higher than the average shift. We also require that words obeying the previous constraints occur more than the average word frequency over the two corpora. Low frequency words are in general less stable (Schnabel et al., 2015), suggesting that shifts for the latter might not be reliable. High frequency words shift globally less (cf. Figure 2), so a higher than average shift could be meaningful.

Figure 6 shows the plot of words that have more or less the same relative frequency in the two newspapers ($-0.1 < gap > 0.1$ and an absolute cumulative frequency higher than average), and we therefore infer that their higher than average shift is mainly due to usage difference. Some comments are provided next to the plot.

These words can be the focus of a dedicated study, and independently of the specific observations that we can make in this context, this method can serve as a way to highlight the hotspot words that deserve attention in a meaning shift study.

4.3 A closer look at nearest neighbours

As a last, more qualitative, analysis, one can inspect how the nearest neighbours of a given word of interest change from one space to the next. In our specific case, we picked a few words (deriving them from the top-down, thus most frequent, and bottom-up selections), and report in Table 2 their top five nearest neighbours in *SpaceR* and in *SpaceRG*. As in most analyses of this kind, one has to rely quite a bit on background and general knowledge to interpret the changes. If we look at “Renzi”, for example, a past Prime Minister from the party close to the newspaper “la Repubblica”, we see that while in $SpaceR$ the top neighbours are all members of his own party, and the party itself (“Pd”), in $SpaceRG$ politicians from other parties (closer to “Il Giornale”) get closer to Renzi, such as Berlusconi and Alfano.

5 Conclusions

We experimented with using embeddings shifts as a tool to study how words are used in two different Italian newspapers. We focused on a pre-selection of high frequency words shared by the two newspapers, and on another set of words which were

[3] Note that this does not correspond exactly to the *gap* measure in Eq. 1 since we are considering the difference between two words rather than the difference in occurrence of the same word in the two corpora.

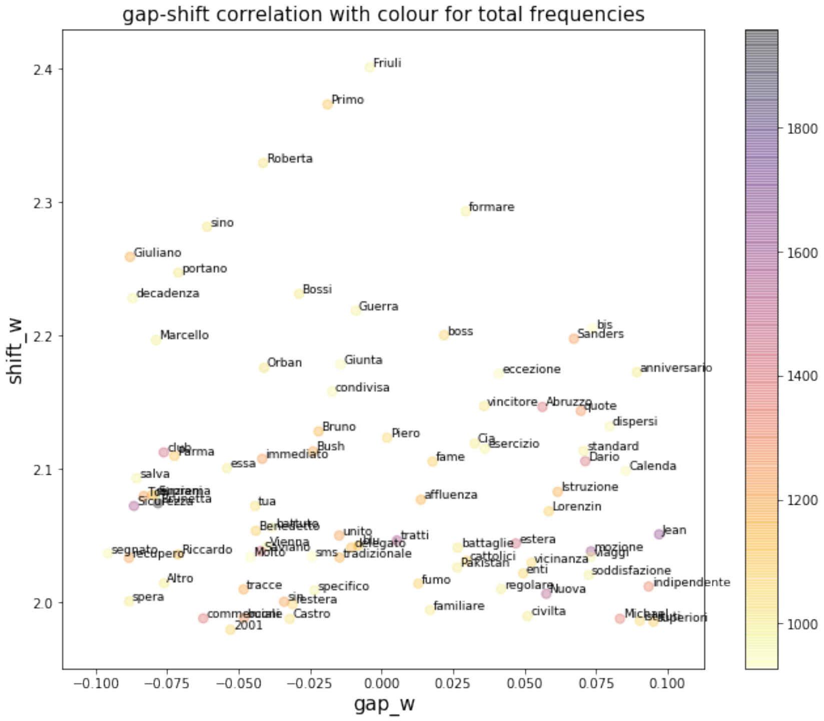

Figure 6: Gap-Shift scatter plot like in Figure 2, zoomed in the gap region -0.1 - 0.1 and shift greater than 1.978 (average shift). Only words with cumulative frequency higher than average frequency are plotted.

Out of the resulting 89 words, 28 are named entities, including politician (Bossi, Lorenzin, Castro, Bush, Saviano, Orban), organisation (CIA, Istruzione[a], Nuova[b]), and location names (Friuli, Vienna, Parma, Pakistan). Common nouns are mostly related to political irksome aspects on which the two newspapers might indeed take diverging positions, such as "sicurezza", "estera", "boss". Other words clearly related to political positions are "renziani", "cattolici", and "tradizionale", the last two being probably part of the 2019 political debate on the so-called traditional family in Italy.

Table 2: A few significant words and their top 5 nearest neighbours in $SpaceR$ and $SpaceRG$.

SpaceR	*SpaceRG*
"migranti" [*en: migrants*]	
barconi [*large boats*] (0.60)	eritrei [*Eritreans*] (0.61)
naufraghi [*castaways*] (0.57)	Lampedusa [] (0.60)
disperati [*wretches*] (0.56)	accoglienza [*hospitality*] (0.59)
barcone [*large boat*] (0.55)	Pozzallo [] (0.58)
carrette [*wrecks*] (0.53)	extracomunitari [*non-European*] (0.57)
"Renzi " [past Prime Minister]	
Orfini [] (0.65)	premier [] (0.60)
Letta [] (0.64)	Nazareno [] (0.59)
Cuperlo [] (0.63)	Berlusconi [] (0.58)
Pd [] (0.62)	Cav [] (0.57)
Bersani [] (0.61)	Alfano [] (0.56)
"politica " [*en: politics*]	
leadership [] (0.65)	tecnocrazia [*technocracy*] (0.60)
logica [*logic*] (0.64)	democrazia [*democracy*] (0.59)
miri [*aspire to*] (0.63)	partitica [*of party*] (0.58)
ambizione [*ambition*] (0.62)	democratica [*democratic*] (0.57)
potentati [*potentates*] (0.61)	legalità [*legality*] (0.56)

highlighted as potentially interesting through a newly proposed methodology which combines observed embeddings shifts and relative and absolute frequency. Most differently used words in the two newspapers are proper nouns of politically active individuals as well as places, and concepts that are highly debated on the political scene.

Beside the present showcase, we believe this methodology can be more in general used to highlight which words might deserve deeper, dedicated analysis when studying meaning change.

One aspect that should be further investigated is the role played by the methodology used for aligning and/or updating the embeddings. As an alternative to what we proposed, one could employ different strategies to manipulate embedding spaces towards highlighting meaning changes. For example, Rodda et al. (2016) exploited Representational Similarity Analysis (Kriegeskorte and Kievit, 2013) to compare embeddings built on different spaces in the context of studying diachronic semantic shifts in ancient Greek. Another interesting approach, still in the context of diachronic meaning change, but applicable to our datasets, was introduced by Hamilton et al. (2016a), who use both a global and a local neighborhood measure of semantic change to disentangle shifts due to cultural changes from purely linguistic ones.

Acknowledgments

We would like to thank the Center for Information Technology of the University of Groningen for providing access to the Peregrine high performance computing cluster.

[a] Most likely a token from the expression "Ministero dell'Istruzione, dell'Università e della Ricerca.

[b] Most likely a token of the bigram "Forza Nuova", an extreme right political movemement.

References

Marco Del Tredici, Malvina Nissim, and Andrea Zaninello. 2016. Tracing metaphors in time through self-distance in vector spaces. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 2116. NIH Public Access.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. 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.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA, June. Association for Computational Linguistics.

Nikolaus Kriegeskorte and Rogier A Kievit. 2013. Representational geometry: integrating cognition, computation, and the brain. *Trends in cognitive sciences*, 17(8):401–412.

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013, 01.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Martina Astrid Rodda, Marco SG Senaldi, and Alessandro Lenci. 2016. Panta rei: Tracking semantic change with distributional semantics in ancient greek. In *CLiC-it/EVALITA*.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.