Analysing Lexical Semantic Change
with Contextualised Word Representations

Mario Giulianelli   **Marco Del Tredici**   **Raquel Fernández**
Institute for Logic, Language and Computation
University of Amsterdam
`{m.giulianelli|m.deltredici|raquel.fernandez}@uva.nl`

Abstract

This paper presents the first unsupervised approach to lexical semantic change that makes use of contextualised word representations. We propose a novel method that exploits the BERT neural language model to obtain representations of word usages, clusters these representations into usage types, and measures change along time with three proposed metrics. We create a new evaluation dataset and show that the model representations and the detected semantic shifts are positively correlated with human judgements. Our extensive qualitative analysis demonstrates that our method captures a variety of synchronic and diachronic linguistic phenomena. We expect our work to inspire further research in this direction.

1 Introduction

In the fourteenth century the words *boy* and *girl* referred respectively to a male servant and a young person of either sex (Oxford English Dictionary). By the fifteenth century a narrower usage had emerged for *girl*, designating exclusively female individuals, whereas by the sixteenth century *boy* had lost its servile connotation and was more broadly used to refer to any male child, becoming the masculine counterpart of *girl* (Bybee, 2015). Word meaning is indeed in constant mutation and, since correct understanding of the meaning of individual words underpins general machine reading comprehension, it has become increasingly relevant for computational linguists to detect and characterise lexical semantic change—e.g., in the form of laws of semantic change (Dubossarsky et al., 2015; Xu and Kemp, 2015; Hamilton et al., 2016)—with the aid of quantitative and reproducible evaluation procedures (Schlechtweg et al., 2018).

Most recent studies have focused on *shift detection*, the task of deciding whether and to what extent the concept evoked by a word has changed

between time periods (e.g., Gulordava and Baroni, 2011; Kim et al., 2014; Kulkarni et al., 2015; Del Tredici et al., 2019; Hamilton et al., 2016; Bamler and Mandt, 2017; Rosenfeld and Erk, 2018). This line of work relies mainly on distributional semantic models, which produce one abstract representation for every word form. However, aggregating all senses of a word into a single representation is particularly problematic for semantic change as word meaning hardly ever shifts directly from one sense to another, but rather typically goes through polysemous stages (Hopper et al., 1991). This limitation has motivated recent work on word sense induction across time periods (Lau et al., 2012; Cook et al., 2014; Mitra et al., 2014; Frermann and Lapata, 2016; Rudolph and Blei, 2018; Hu et al., 2019). Word senses, however, have shortcomings themselves as they are a discretisation of word meaning, which is continuous in nature and modulated by context to convey ad-hoc interpretations (Brugman, 1988; Kilgarriff, 1997; Paradis, 2011).

In this work, we propose a usage-based approach to lexical semantic change, where sentential context modulates lexical meaning "on the fly" (Ludlow, 2014). We present a novel method that (1) exploits a pre-trained neural language model (BERT; Devlin et al., 2019) to obtain contextualised representations for every occurrence of a word of interest, (2) clusters these representations into *usage types*, and (3) measures change along time. More concretely, we make the following contributions:

- We present the first unsupervised approach to lexical semantic change that makes use of state-of-the-art contextualised word representations.

- We propose several metrics to measure semantic change with this type of representation. Our code is available at `https://github.com/glnmario/cwr4lsc`.

- We create a new evaluation dataset of human sim-

ilarity judgements on more than 3K word usage pairs across different time periods, available at `https://doi.org/10.5281/zenodo.3773250`.

- We show that both the model representations and the detected semantic shifts are positively correlated with human intuitions.

- Through in-depth qualitative analysis, we show that the proposed approach captures synchronic phenomena such as word senses and syntactic functions, literal and metaphorical usage, as well as diachronic linguistic processes related to narrowing and broadening of meaning across time.

Overall, our study demonstrates the potential of using contextualised word representations for modelling and analysing lexical semantic change and opens the door to further work in this direction.

2   Related Work

Semantic change modelling   Lexical semantic change models build on the assumption that meaning change results in the modification of a word's linguistic distribution. In particular, with the exception of a few methods based on word frequencies and parts of speech (Michel et al., 2011; Kulkarni et al., 2015), lexical semantic change detection has been addressed following two main approaches: *form-based* and *sense-based* (for an overview, see Kutuzov et al., 2018; Tang, 2018).

In *form-based* approaches independent models are trained on the time intervals of a diachronic corpus and the distance between representations of the same word in different intervals is used as a semantic change score (Gulordava and Baroni, 2011; Kulkarni et al., 2015). Representational coherence between word vectors across different periods can be guaranteed by incremental training procedures (Kim et al., 2014) as well as by post hoc alignment of semantic spaces (Hamilton et al., 2016). More recent methods capture diachronic word usage by learning dynamic word embeddings that vary as a function of time (Bamler and Mandt, 2017; Rosenfeld and Erk, 2018; Rudolph and Blei, 2018). Form-based models depend on a strong simplification: that a single representation is sufficient to model the different usages of a word.

Time-dependent representations are also created in *sense-based* approaches: in this case word meaning is encoded as a distribution over word senses. Several Bayesian models of sense change have been proposed (Wijaya and Yeniterzi, 2011; Lau

et al., 2012, 2014; Cook et al., 2014). Among these is the recent SCAN model (Frermann and Lapata, 2016), which represents (1) the meaning of a word in a time interval as a multinomial distribution over word senses and (2) word senses as probability distributions over the vocabulary. The main limitation of sense-based models is that they rely on a bag-of-words representation of context. Furthermore, many of these models keep the number of senses constant across time intervals and require this number to be manually set in advance.

Unsupervised approaches have been proposed that do not rely on a fixed number of senses. For example, the method for novel sense identification by Mitra et al. (2015) represents senses as clusters of short dependency-labelled contexts. Like ours, this method analyses word forms within the grammatical structures they appear. However, it requires syntactically parsed diachronic corpora and focuses exclusively on nouns. None of these restrictions limit our proposed approach, which leverages neural contextualised word representations.

Contextualised word representations   Several approaches to context-sensitive word representations have been proposed in the past. Schütze (1998) introduced a clustering-based disambiguation algorithm for word usage vectors, Erk and Padó (2008) proposed creating multiple vectors for the same word and Erk and Padó (2010) proposed to directly learn usage-specific representations based on the set of exemplary contexts within which the target word occurs.

Recently, neural contextualised word representations have gained widespread use in NLP, thanks to deep learning models which learn usage-dependent representations while optimising tasks such as machine translation (CoVe; McCann et al., 2017) and language modelling (Dai and Le, 2015, ULMFiT; Howard and Ruder, 2018, ELMo; Peters et al., 2018, GPT; Radford et al., 2018, 2019, BERT; Devlin et al., 2019). State-of-the-art language models typically use stacked attention layers (Vaswani et al., 2017), they are pre-trained on a very large amount of textual data, and they can be fine-tuned for specific downstream tasks (Howard and Ruder, 2018; Radford et al., 2019; Devlin et al., 2019).

Contextualised representations have been shown to encode lexical meaning dynamically, reaching high accuracy on, e.g., the binary usage similarity judgements of the WiC evaluation set (Pilehvar and Camacho-Collados, 2019), performing on a

par with state-of-the-art word sense disambiguation models (Wiedemann et al., 2019), and proving useful for the supervised derivation of time-specific sense representation (Hu et al., 2019). In this work, we investigate the potential of contextualised word representations to detect and analyse lexical semantic change, without any lexicographic supervision.

3 Method: A Usage-based Approach to Lexical Semantic Change

We introduce a usage-based approach to lexical semantic change analysis which relies on contextualised representations of unique word occurrences (*usage representations*). First, given a diachronic corpus and a list of words of interest, we use the BERT language model (Devlin et al., 2019) to compute usage representations for each occurrence of these words. Then, we cluster all the usage representations collected for a given word into an automatically determined number of partitions (*usage types*) and organise them along the temporal axis. Finally, we propose three metrics to quantify the degree of change undergone by a word.

3.1 Language Model

We produce usage representations using the BERT language model (Devlin et al., 2019), a multi-layer bidirectional Transformer encoder trained on masked token prediction and next sentence prediction, on the BooksCorpus (800M words) (Zhu et al., 2015) and on English text passages extracted from Wikipedia (2,500M words). There are two versions of BERT. For space and time efficiency, we use the smaller *base-uncased* version, with 12 layers, 768 hidden dimensions, and 110M parameters.[1]

3.2 Usage Representations

Given a word of interest w and a context of occurrence $s = (v_1, ..., v_i, ..., v_n)$ with $w = v_i$, we extract the activations of all of BERT's hidden layers for sentence position i and sum them dimension-wise. We use addition because neither concatenation nor selecting a subset of the layers produced notable differences in the relative geometric distance between word representations.

The set of N usage representations for w in a given corpus can be expressed as the usage matrix $\mathbf{U}_w = (\mathbf{w}_1, \dots, \mathbf{w}_N)$. For each usage representation in the usage matrix $\mathbf{U}_w$, we store the context of

(a) PCA visualisation of the usage representations.

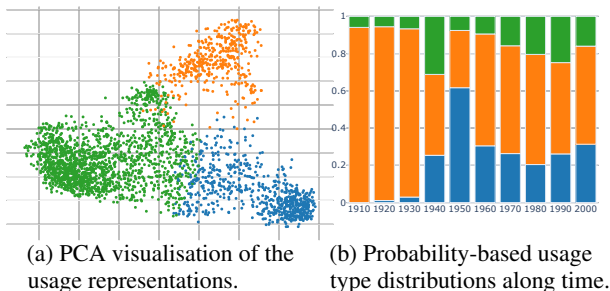(b) Probability-based usage type distributions along time.

Figure 1: Usage representations and usage type distributions generated with occurrences of the word *atom* in COHA (Davies, 2012). Colours encode usage types.

occurrence (a 128-token window around the target word) as well as a temporal label $\mathbf{t}_w$ indicating the time interval of the usage.

3.3 Usage Types

Once we have obtained a word-specific matrix of usage vectors $\mathbf{U}_w$, we standardise it and cluster its entries using K-Means.[2] This step partitions usage representations into clusters of similar usages of the same word, or *usage types* (see Figure 1a), and thus it is directly related to automatic word sense discrimination (Schütze, 1998; Pantel and Lin, 2002; Manandhar et al., 2010; Navigli and Vannella, 2013, among others).

For each word independently, we automatically select the number of clusters K that maximises the silhouette score (Rousseeuw, 1987), a metric of cluster quality which favours intra-cluster coherence and penalises inter-cluster similarity, without the need for gold labels. For each value of K, we execute 10 iterations of Expectation Maximization to alleviate the influence of different initialisation values (Arthur and Vassilvitskii, 2007). The final clustering for a given K is the one that yields the minimal *distortion* value across the 10 runs, i.e., the minimal sum of squared distances of each data point from its closest centroid. We experiment with $K \in [2, 10]$. We choose the range $[2, 10]$ heuristically: we forgo $K = 1$ as K-Means and the silhouette score are ill-defined for this case, while keeping the number of possible clusters manageable computationally. This excludes the possibility that a word has a single usage type. Alternatively, we could use a measure of intra-cluster dispersion for $K = 1$, and consider a word monosemous if its dispersion value is below a threshold d (if the dispersion is higher than d, we would discard $K = 1$

---

[1]We rely on *Hugging Face*'s implementation of BERT (available at https://github.com/huggingface/transformers).

[2]Other clustering methods are also possible. For this first study, we choose the widely used K-Means (*scikit-learn*).

and use the silhouette score to find the best $K \geq 2$). There also exist clustering methods that select the optimal K automatically, e.g. DBSCAN or Affinity Propagation (Martinc et al., 2020). They nevertheless require method-specific parameter choices which indirectly determine the number of clusters.

By counting the number of occurrences of each usage type k in a given time interval t (we refer to this count as $freq(k, t)$), we obtain frequency distributions $\mathbf{f}_w^t$ for each interval under scrutiny:

$$\mathbf{f}_w^t \in \mathbb{N}^{K_w} : \mathbf{f}_w^t[k] = freq(k, t) \quad k \in [1, K_w] \quad (1)$$

When normalised, frequency distributions can be interpreted as probability distributions over usage types $\mathbf{u}_w^t : \mathbf{u}_w^t[k] = \frac{1}{N_t}\mathbf{f}_w^t[k]$. Figure 1b illustrates the result of this process.

3.4 Quantifying Semantic Change

We propose three metrics for the automatic quantification of lexical semantic change using contextualised word representations. The first two (*entropy difference* and *Jensen-Shannon divergence*) are known metrics for comparing probability distributions. In our approach, we apply them to measure variations in the relative prominence of coexisting usage types. We conjecture that these kinds of metric can help detect semantic change processes that, e.g., lead to broadening or narrowing (i.e., to increase or decrease, respectively, in the number or relative distribution of usage types).

The third metric (*average pairwise distance*) only requires a usage matrix $\mathbf{U}_w$ and the temporal labels $\mathbf{t}_w$ (Section 3.2). Since it does not rely on usage type distributions, it is not sensitive to possible errors stemming from the clustering process.

Entropy difference (ED) We propose measuring the uncertainty (e.g., due to polysemy) in the interpretation of a word w in interval t using the normalised entropy of its usage distribution $\mathbf{u}_w^t$:

$$\eta(\mathbf{u}_w^t) = \log_{K_w} \left(\prod_{k=1}^{K_w} \mathbf{u}_w^t[k]^{-\mathbf{u}_w^t[k]} \right) \quad (2)$$

To quantify how uncertainty over possible interpretations varies across time intervals, we compute the difference in entropy between the two usage type distributions in these intervals: $\text{ED}(\mathbf{u}_w^t, \mathbf{u}_w^{t'}) = \eta(\mathbf{u}_w^{t'}) - \eta(\mathbf{u}_w^t)$. We expect high ED values to signal the broadening of a word's interpretation and negative values to indicate narrowing.

Jensen-Shannon divergence (JSD) The second metric takes into account not only variations in the size of usage type clusters but also *which clusters* have grown or shrunk. It is the Jensen-Shannon divergence (Lin, 1991) between usage type distributions:

$$\text{JSD}(\mathbf{u}_w^t, \mathbf{u}_w^{t'}) = \text{H} \left(\frac{1}{2} \left(\mathbf{u}_w^t + \mathbf{u}_w^{t'} \right) \right) \\ - \frac{1}{2} \left(\text{H} \left(\mathbf{u}_w^t \right) - \text{H} \left(\mathbf{u}_w^{t'} \right) \right) \quad (3)$$

where H is the Boltzmann-Gibbs-Shannon entropy. Very dissimilar usage distributions yield high JSD whereas low JSD values indicate that the proportions of usage types barely change across periods.

Average pairwise distance (APD) While the previous two metrics rely on usage type distributions, it is also possible to quantify change bypassing the clustering step into usage types, e.g. by calculating the average pairwise distance between usage representations in different periods t and t':

$$\text{APD}(\mathbf{U}_w^t, \mathbf{U}_w^{t'}) = \frac{1}{N^t \cdot N^{t'}} \sum_{\mathbf{x}_i \in \mathbf{U}_w^t, \ \mathbf{x}_j \in \mathbf{U}_w^{t'}} d(\mathbf{x}_i, \mathbf{x}_j) \quad (4)$$

where $\mathbf{U}_w^t$ is a usage matrix constructed with occurrences of w only in interval t. We experiment with cosine, Euclidean, and Canberra distance.

Generalisation to multiple time intervals The presented metrics quantify semantic change across pairs of temporal intervals (t, t'). When more than two intervals are available, we measure change across all contiguous intervals $(m(\mathbf{U}_w^t, \mathbf{U}_w^{t+1})$, where m is one of the metrics), and collect these values into vectors. We then transform each vector into a scalar change score by computing the vector's mean and maximum values.[3] Whereas the *mean* is indicative of semantic change across the entire period under consideration, the *max* pinpoints the pair of successive intervals where the strongest shift has occurred.

---

[3]The Jensen-Shannon divergence can also be measured with respect to $T > 2$ probability distributions (Ré and Azad, 2014): $\text{JSD} \left(\mathbf{u}_w^1, \ldots, \mathbf{u}_w^T \right) = \text{H} \left(\frac{1}{T} \sum_{i=1}^{T} \mathbf{u}_w^i \right) - \frac{1}{T} \sum_{i=1}^{T} \text{H} \left(\mathbf{u}_w^i \right)$. However, this definition of the JSD is insensitive to the order of the temporal intervals and yields lower correlation with human semantic change ratings (cfr. Section 5.2) than the pairwise metrics.

4 Data

We examine word usages in a large diachronic corpus of English, the Corpus of Historical American English (COHA, Davies, 2012), which covers two centuries (1810–2009) of language use and includes a variety of genres, from fiction to newspapers and popular magazines, among others. In this study, we focus on texts written between 1910 and 2009, for which a minimum of 21M words per decade is available, and discard previous decades, where data are less balanced per decade.

We use the 100 words annotated with semantic shift scores by Gulordava and Baroni (2011) as our target words. These scores are human judgements collected by asking five annotators to quantify the degree of semantic change undertaken by each word (shown out of context) from the 1960's to the 1990's. We exclude *extracellular* as in COHA this word only appears in three decades; all other words appear in at least 8 decades, with a minimum and maximum frequency of 191 and 108,796, respectively. We refer to the resulting set of 99 words and corresponding shift scores as the 'GEMS dataset' or the 'GEMS words', as appropriate.

We collect a contextualised representation for each occurrence of these words in the second century of COHA, using BERT as described in Section 3.2. This results in a large set of usage representations, ∼1.3M in total, which we cluster into usage types using K-Means and silhouette coefficients (Section 3.3). We use these usage representations and usage types in the evaluation and the analyses offered in the remaining of the paper.

5 Correlation with Human Judgements

Before using our proposed method to analyse language change, we assess how its key components compare with human judgements. We test whether the clustering into usage types reflects human similarity judgements (Section 5.1) and to what extent the degree of change computed with our metrics correlates with shift scores provided by humans (Section 5.2).

5.1 Evaluation of Usage Types

The clustering of contextualised representations into usage types is one of the main steps in our method (see Section 3.3). It relies on the similarity values between pairs of usage representations created by the language model. To quantitatively evaluate the quality of these similarity values (and thus,

by extension, the quality of usage representations and usage types), we compare them to similarity judgements by human raters.

New dataset of similarity judgements We create a new evaluation dataset, following the annotation approach of Erk et al. (2009, 2013) for rating pairs of usages of the same word. Since we need to collect human judgements for pairs of usages, annotating the entire GEMS dataset would be extremely costly and time consuming. Therefore, to limit the scope of the annotation, we select a subset of words. For each shift score value s in the GEMS dataset, we sample a word uniformly at random from the words annotated with s. This results in 16 words. To ensure that our selection of usages is sufficiently varied, for each of these words, we sample five usages from each of their usage types (the number of usage types is word-specific) along different time intervals, one usage per 20-year period over the century. All possible pairwise combinations are generated for each target word, resulting in a total of 3,285 usage pairs.

We use the crowdsourcing platform Figure Eight[4] to collect five similarity judgements for each of these usage pairs. Annotators are shown pairs of usages of the same word: each usage shows the target word in its sentence, together with the previous and the following sentences (67 tokens on average). Annotators are asked to assign a similarity score on a 4-point scale, ranging from *unrelated* to *identical*, as defined by Brown (2008) and used e.g., by Schlechtweg et al. (2018).[5] A total of 380 annotators participated in the task. The inter-rater agreement, measured as the average pairwise Spearman's correlation between common annotation subsets, is 0.59. This is in line with previous approaches such as Schlechtweg et al. (2018), who report agreement scores between 0.57 and 0.68.

Results To obtain a single human similarity judgement per usage pair, we average the scores given by five annotators. We encode all averaged human similarity judgements for a given word in a square matrix. We then compute similarity scores over pairs of usage vectors output by BERT[6] to

---

[4] https://www.figure-eight.com, recently acquired by Appen (https://appen.com).

[5] The full instructions with examples given to the annotators are available in Appendix A.1.

[6] For this evaluation, BERT is given the same variable-size context as the human annotators. Vector similarity values are computed as the inverse of Euclidean distance, because K-means relies on this metric for cluster assignments.

obtain analogous matrices per word and measure Spearman's rank correlation between the human- and the machine-generated matrices using the Mantel test (Mantel, 1967).

We observe a significant ($p < 0.05$) positive correlation for 10 out of 16 words, with ρ coefficients ranging from 0.13 to 0.45.[7] This is an encouraging result, which indicates that BERT's word representations and similarity scores (as well as our clustering methods which build on them) correlate, to a substantial extent, with human similarity judgements. We take this to provide a promising empirical basis for our approach.

5.2 Evaluation of Semantic Change Scores

We now quantitatively assess the semantic change scores yielded by the metrics described in Section 3.4 when applied to BERT usage representations and the usage types created with our approach. We do so by comparing them to the human shift scores in the GEMS dataset. For consistency with this dataset, which quantifies change from the 1960's to the 1990's as explained in Section 4, we only consider these four decades when calculating our scores. Using each of the metrics on representations from these time intervals, we assign a semantic change score to all the GEMS words. We then compute Spearman's rank correlation between the automatically generated change scores and the gold standard shift values.

Results Table 1 shows the Spearman's correlation coefficients obtained using our metrics, together with a frequency baseline (the difference between the normalised frequency of a word in the 1960's and in the 1990's). The three proposed metrics yield significant positive correlations. This is again a very encouraging result regarding the potential of contextualised word representations for capturing lexical semantic change.

As a reference, we report the correlation coefficients with respect to GEMS shift scores documented by the authors of two alternative approaches: the count-based model by Gulordava and Baroni (2011) themselves (trained on two time slices from the Google Books corpus with texts from the 1960's and the 1990's) and the sense-based SCAN model by Frermann and Lapata (2016) (trained on the DATE corpus with texts from the 1960's through the 1990's).[8]

Frequency difference	0.068
Entropy difference (*max*)	0.278
Jensen-Shannon divergence (*max*)	0.276
Average pairwise distance (*Euclidean*, *max*)	0.285
Gulordava and Baroni (2011)	0.386
Frermann and Lapata (2016)	0.377

Table 1: Spearman's ρ correlation coefficients between the gold standard scores in the GEMS dataset and the change scores assigned by our three metrics and a relative frequency baseline. For reference, correlation coefficients reported by previous works using different approaches are also given. All correlations are significant ($p < 0.05$) except for the frequency difference baseline.

For all our metrics, the *max* across the four time intervals—i.e., identifying the pair of successive intervals where the strongest shift has occurred (cfr. end of Section 3.4)—is the best performing aggregation strategy. Table 1 only shows values obtained with *max* and Euclidean distance for APD, as they are the best-performing options.

It is interesting to observe that APD can prove as informative as JSD and ED, although it does not depend on the clustering of word occurrences into usage types. Yet, computing usage types offers a powerful tool for analysing lexical change, as we will see in the next section.

6 Analysis

In this section, we provide an in-depth qualitative analysis of the linguistic properties that define usage types and the kinds of lexical semantic change we observe. More quantitative methods (such as taking the top n words with highest JSD, APD and ED and checking, e.g., how many cases of broadening each metric captures) are difficult to operationalise (Tang et al., 2016) because there exist no well-established formal notions of semantic change types in the linguistic literature. To carry out this analysis, for each GEMS word, we identify the most representative usages in a given usage type cluster by selecting the five closest vectors to the cluster centroid, and take the five corresponding sentences as usage examples.

6.1 What do Usage Types Capture?

We first leave the temporal variable aside and present a synchronic analysis of usage types. Our

---

[7]Scores per target word are given in Appendix A.2.

[8]Gulordava and Baroni (2011) report Pearson correlation.

However, to allow for direct comparison, Frermann and Lapata (2016) computed Spearman correlation for that work (see their footnote 7), which is the value we report.

goal is to assess the interpretability and internal coherence of the obtained usage clusters.

We observe that usage types can discriminate between underlying senses of polysemous (and homonymous) words, between literal and figurative usages, and between usages that fulfil different syntactic roles; plus they can single out phrasal collocations as well as named entities.

Polysemy and homonymy Distinctions often occur between underlying senses of polysemous and homonymous words. For example, the vectors collected for the polysemous word *curious* are grouped together into two usage types, depending on whether *curious* is used to describe something that excites attention as odd, novel, or unexpected ('a wonderful and *curious* and unbelievable story') or rather to describe someone who is marked by a desire to investigate and learn ('*curious* and amazed and innocent'). The same happens for the homonymous usages of the word *coach*, for instance, which can denote vehicles as well as instructors (see Figure 2a for a diachronic view of the usage types).

Metaphor and metonymy In several cases, literal and metaphorical usages are also separated. For example, occurrences of *curtain* are clustered into four usage types (Figure 2c): two of these correspond to a literal interpretation of the word as a hanging piece of cloth ('*curtain*less windows', 'pulled the *curtain* closed') whereas the other two indicate metaphorical interpretations of *curtain* as any barrier that excludes the free exchange of information or communication ('the *curtain* on the legal war is being raised'). Similarly, we obtain two usage types for *sphere*: one for literal usages that denote a round solid figure ('the *sphere* of the moon'), and the other for metaphorical interpretations of the word as an area of knowledge or activity ('a certain *sphere* of autonomy') as well as metonymical usages that refer to the planet Earth ('land and peoples on the top half of the *sphere*').

Syntactic roles and argument structure Further distinctions are observed between word usages that fulfil a different syntactic functionality: not only is part-of-speech ambiguity detected (e.g., 'the *cost*-tapered average tariff' vs. '*cost* less to make') but contextualised representations also capture regularities in syntactic argument structures. For example, usages of *refuse* are clustered into nominal usages ('society's emotional *refuse*', 'the amount of *refuse*'), verbal transitive and intransi-

tive usages ('fall, give up, *refuse*, kick'), as well as verbal usages with infinitive complementation ('*refuse* to go', '*refuse* for the present to sign a treaty').

Collocations and named entities Specific clusters are also assigned to lexical items that are parts of phrasal collocations (e.g., 'iron *curtain*') or of named entities ('alexander graham *bell*' vs. '*bell*-like whistle').

Other distinctions Some distinctions are interpretable but unexpected. As an example, the word *doubt* does not show the default noun-verb separation but rather a distinction between usages in affirmative contexts ('there is still *doubt*', 'the benefit of the *doubt*') and in negative contexts ('there is not a bit of *doubt*', 'beyond a reasonable *doubt*').

Observed errors For some words, we find that usages which appear to be identical are separated into different usage types. In a handful of cases, this seems due to the setup we have used for experimentation, which sets the minimum number of clusters to 2 (see Section 3.3). This leads to distinct usage types for words such as *maybe*, for which a single type is expected. In other cases, a given interpretation is not identified as an independent type, and its usages appear in different clusters. This holds, for example, for the word *tenure*, whose usages in phrases such as '*tenure*-track faculty position' are present in two distinct usage types (see Figure 2b).

Finally, we see that in some cases a usage type ends up including two interpretations which arguably should have been distinguished. For example, two of the usage types identified for *address* are interpretable and coherent: one includes usages in the sense of formal speech and the other one includes verbal usages. The third usage type, however, includes a mix of nominal usages of the word as in 'disrespectful manners or *address*' as well as in 'network *address*'.

6.2 What Kinds of Change are Observed?

Here we consider usage types diachronically. Different kinds of change, driven by cultural and technological innovation as well as by historical events, emerge from a qualitative inspection of usage distributions along the temporal dimension. We describe the most prominent kinds—narrowing and broadening, including metaphorisation—and discuss the extent to which our metrics are able to detect them.
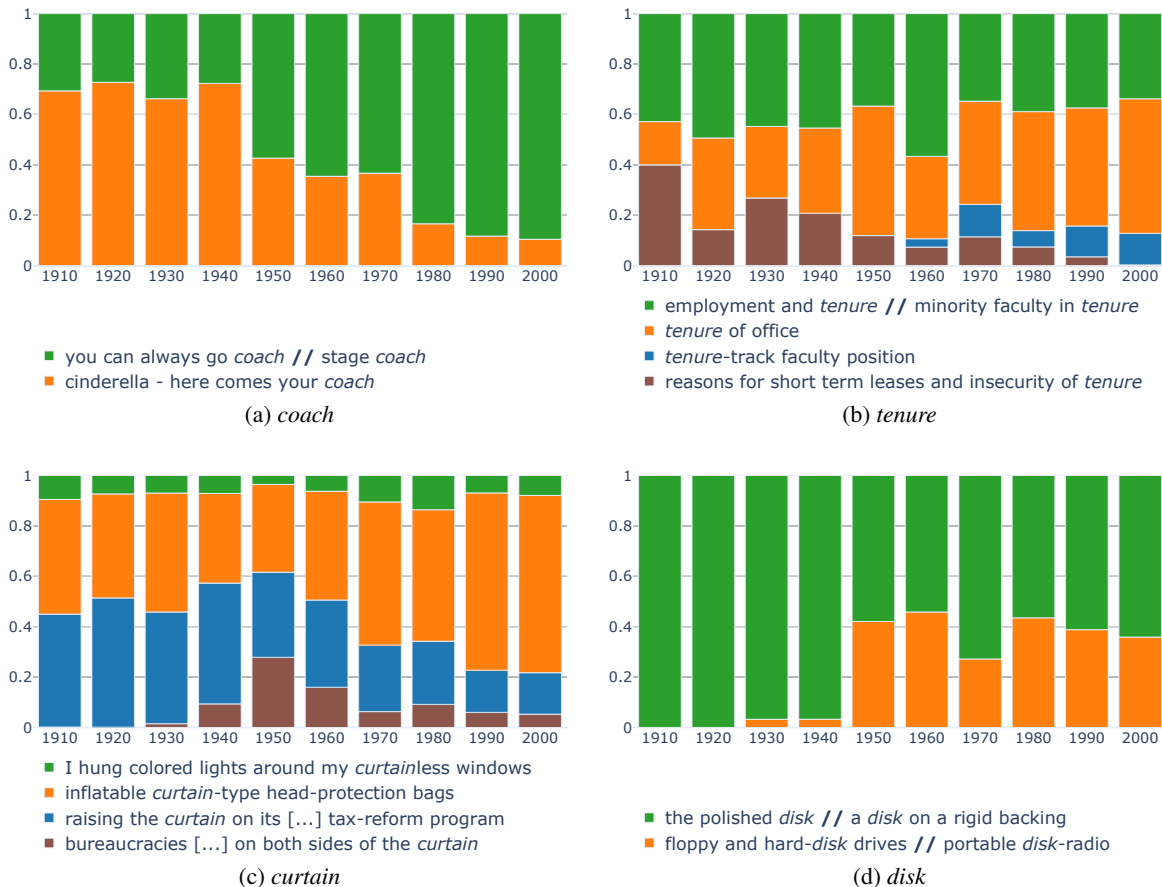
Figure 2: Evolution of usage type distributions in the period 1910–2009, generated with occurrences of *coach*, *tenure*, *curtain* and *disk* in COHA (Davies, 2012). The legends show sample usages per identified usage type.

Narrowing Examination of the dynamics of usage distributions allows us to see that for a few words certain usage types disappear or become less common over time (i.e., the interpretation of the word becomes 'narrower', less varied). This is the case, for example, for *coach*, where the frequency decrease of a usage type is gradual and caused by technological evolution (see Figure 2a).

Negative mean ED (see Section 3.4) reliably indicates this kind of narrowing. Indeed *coach* is assigned one of the lowest ED score among the GEMS words. In contrast, ED fails to detect the obsolescence of a usage type when new usage types emerge simultaneously (since this may lead to no entropy reduction). This is the case, e.g., of *tenure*. The usage type capturing *tenure* of a landed property becomes obsolete; however, we obtain a positive mean ED caused by the appearance of a new usage type (the third type in Figure 2b).

Broadening For a substantial amount of words, we observe the emergence of new usage types (i.e., a 'broadening' of their use). This may be due to

technological advances as well as to specific historical events. As an example, Figure 2d shows how, starting from the 1950's and as a result of technological innovation, the word *disk* starts to be used to denote also optical disks while beforehand it referred only to generic flat circular objects.

A special kind of broadening is metaphorisation. As mentioned in Section 6.1, the usage types for the word *curtain* include metaphorical interpretations. Figure 2c allows us to see when the metaphorical meaning related to the historically charged expression *iron curtain* is acquired. This novel usage type is related to a specific historical period: it emerges between the 1930's and the 1940's, reaches its peak in the 1950's, and remains stably low in frequency starting from the 1970's.

The metrics that best capture broadening are JSD and APD—e.g., *disk* is assigned a high semantic change score by both metrics. Yet, sometimes these metrics generate different score rankings. For example, *curtain* yields a rather low APD score due to the low relative frequency of the novel usage (Figure 2c). In contrast, even though the novel us-

age type is not very prominent in some decades, JSD can still discriminate it and measure its development. On the other hand, the word *address*, for which we also observe broadening, is assigned a low score by JSD due to the errors in its usage type assignments pointed out in Section 6.1. As APD does not rely on usage types, it is not affected by this issue and does indeed assign a high change score to the word.

Finally, although our metrics help us identify the broadening of a word's meaning, they cannot capture the type of broadening (i.e., the nature of the emerging interpretations). Detecting metaphorisation, for example, may require inter-cluster comparisons to identify a metaphor's source and target usage types, which we leave to future work.

7 Conclusion

We have introduced a novel approach to the analysis of lexical semantic change. To our knowledge, this is the first work that tackles this problem using neural contextualised word representations and no lexicographic supervision. We have shown that the representations and the detected semantic shifts are aligned to human interpretation, and presented a new dataset of human similarity judgements which can be used to measure said alignment. Finally, through extensive qualitative analysis, we have demonstrated that our method allows us to capture a variety of synchronic and diachronic linguistic phenomena.

Our approach offers several advantages over previous methods: (1) it does not rely on a fixed number of word senses, (2) it captures morphosyntactic properties of word usage, and (3) it offers a more effective interpretation of lexical meaning by enabling the inspection of particular example sentences. In recent work, we have experimented with alternative ways of obtaining usage representations (using a different language model, fine-tuning, and various layer selection strategies) and we have obtained very promising results in detecting semantic change across four languages (Kutuzov and Giulianelli, 2020). In the future, we plan to investigate whether usage representations can provide an even finer grained account of lexical meaning and its dynamics, e.g., to automatically discriminate between different types of meaning change. We expect our work to inspire further analyses of variation and change which exploit the expressiveness of contextualised word representations.

References

David Arthur and Sergei Vassilvitskii. 2007. `k-means++`: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics.

Robert Bamler and Stephan Mandt. 2017. Dynamic Word Embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 380–389. JMLR.org.

Susan Windisch Brown. 2008. Choosing Sense Distinctions for WSD: Psycholinguistic Evidence. In *Proceedings of ACL-08: HLT, Short Papers*, pages 249–252, Columbus, Ohio. Association for Computational Linguistics.

Claudia Marlea Brugman. 1988. *The Story of Over: Polysemy, Semantics, and the Structure of the Lexicon*. Garland, New York.

Joan Bybee. 2015. *Language Change*. Cambridge University Press.

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel Word-Sense Identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1635.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised Sequence Learning. In *Advances in Neural Information Processing Systems*, pages 3079–3087.

Mark Davies. 2012. Expanding Horizons in Historical Linguistics with the 400-Million Word Corpus of Historical American English. *Corpora*, 7(2):121–157.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-Term Meaning Shift: A Distributional Exploration. In *Proceedings of NAACL-HLT 2019 (Annual Conference of the North American Chapter of the Association for Computational Linguistics)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Haim Dubossarsky, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. 2015. A Bottom Up Approach to Category Mapping and Meaning Change. In *Word Structure and Word Usage. Proceedings of the NetWordS Final Conference*, pages 66–70.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on Word Senses and Word Usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 10–18, Suntec, Singapore. Association for Computational Linguistics.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring Word Meaning in Context. *Computational Linguistics*, 39(3):511–554.

Katrin Erk and Sebastian Padó. 2008. A Structured Vector Space Model for Word Meaning in Context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 897–906.

Katrin Erk and Sebastian Padó. 2010. Exemplar-Based Models for Word Meaning in Context. In *Proceedings of the ACL 2010 Conference (Short Papers)*, pages 92–97.

Lea Frermann and Mirella Lapata. 2016. A Bayesian Model of Diachronic Meaning Change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Mario Giulianelli. 2019. Lexical Semantic Change Analysis with Contextualised Word Representations. Master's thesis, University of Amsterdam, July.

Kristina Gulordava and Marco Baroni. 2011. A Distributional Similarity Approach to the Detection of Semantic Change in the Google Books Ngram Corpus. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pages 67–71.

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.

Paul J Hopper et al. 1991. On Some Principles of Grammaticization. *Approaches to Grammaticalization*, 1:17–35.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

Adam Kilgarriff. 1997. I Don't Believe in Word Senses. *Computers and the Humanities*, 31(2):91–113.

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.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically Significant Detection of Linguistic Change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection. Forthcoming.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic Word Embeddings and Semantic Shifts: A Survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397.

Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. 2014. Learning Word Sense Distributions, Detecting Unattested Senses and Identifying Novel Senses Using Topic Models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 259–270.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word Sense Induction for Novel Sense Detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601. Association for Computational Linguistics.

Jianhua Lin. 1991. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Peter Ludlow. 2014. *Living Words: Meaning Underdetermination and the Dynamic Lexicon*. OUP Oxford.

Suresh Manandhar, Ioannis P Klapaftis, Dmitriy Dligach, and Sameer S Pradhan. 2010. SemEval-2010 Task 14: Word Sense Induction & Disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68. Association for Computational Linguistics.

Nathan Mantel. 1967. The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, 27(2):209–220.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020. Capturing Evolution in Word Usage: Just Add More Clusters? In *Companion Proceedings of the International World Wide Web Conference*, pages 20–24.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in Translation: Contextualized Word Vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182.

Sunny Mitra, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2015. An Automatic Approach to Identify Word Sense Changes in Text Media across Timescales. *Natural Language Engineering*, 21(5):773–798.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That's Sick Dude! Automatic Identification of Word Sense Change across Different Timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1020–1029.

Roberto Navigli and Daniele Vannella. 2013. SemEval-2013 Task 11: Word Sense Induction and Disambiguation within an End-User Application. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 193–201.

Patrick Pantel and Dekang Lin. 2002. Discovering Word Senses from Text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, page 613–619, New York, NY, USA. Association for Computing Machinery.

Carita Paradis. 2011. Metonymization: A Key Mechanism in Semantic Change. *Defining Metonymy in Cognitive Linguistics: Towards a Consensus View*, pages 61–98.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-training. Technical report, OpenAI.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

Miguel A Ré and Rajeev K Azad. 2014. Generalization of Entropy Based Divergence Measures for Symbolic Sequence Analysis. *PloS one*, 9(4):e93532.

Alex Rosenfeld and Katrin Erk. 2018. Deep Neural Models of Semantic Shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 474–484.

Peter J. Rousseeuw. 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for Language evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1003–1011. International World Wide Web Conferences Steering Committee.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 169–174.

Hinrich Schütze. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–123.

Xuri Tang. 2018. A State-of-the-Art of Semantic Change computation. *Natural Language Engineering*, 24(5):649–676.

Xuri Tang, Weiguang Qu, and Xiaohe Chen. 2016. Semantic Change Computation: A Successive Approach. *World Wide Web*, 19(3):375–415.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019*, Erlangen, Germany.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding Semantic Change of Words over Centuries. In *Proceedings of the 2011 International Workshop on Detecting and Exploiting Cultural Diversity on the Social Web*, pages 35–40. ACM.

Yang Xu and Charles Kemp. 2015. A Computational Evaluation of Two Laws of Semantic Change. In *CogSci*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19–27.

A Appendix

This appendix includes supplementary materials related to Section 5.1.

A.1 New Dataset of Similarity Judgements

Obtaining usage pairs For each of our 16 target words, we sample five usages from each of their usage types, one for every 20-year period in the last century of COHA. When a usage type does not occur in a time interval, we uniformly sample an interval from those that do contain occurrences of that usage type. All possible pairwise combinations (without replacement) are generated for each target word, resulting in a total of 3,285 usage pairs.

Crowdsourced annotation We use the crowdsourcing platform Figure Eight (since then acquired by Appen[9]) to collect five similarity judgements for each of these usage pairs. To control the quality of the similarity judgements, we select Figure Eight workers from the pool of most experienced contributors, we require them to be native English

---

[9] https://appen.com

speakers and to have completed a test quiz consisting of 10 similarity judgements. For this purpose, 170 usage pairs were manually annotated by the first author with 1 to 3 acceptable labels. The compensation scheme for the raters is based on an average wage of 10 USD per hour.

Figures 4 and 5 (on the next pages) show the full instructions given to the annotators and Figure 3 illustrates a single annotation item.

---

federal

Please read carefully the following two sentences where the word [[federal]] occurs:

- robert m . hitchcock , who prosecuted the amerasia case in 1945 , testified today that he had been gravely handicapped because the government ' s best evidence had been produced by illegal seizures by [[federal]] agents . the prosecution , he asserted , was in fact fortunate under the circumstances to have done as well as it did .
- there should be such a fire every saturday afternoon at the same time with the same actual damage . this time it was the records and documents of the [[federal]] trade commision , said to be " priceless . " also the reels of official motion pictures of historical or technical value .

How similar are the two occurrences of [[federal]]? (required)
- 1. Unrelated
- 2. Distantly related
- 3. Closely related
- 4. Identical
- Cannot decide (please use this option as little as possible)

Figure 3: An annotation item on the Figure Eight crowdsourcing platform.

A.2 Correlation Results

We measure Spearman's rank correlation between human- and machine-generated usage similarity matrices using the Mantel test and observe a significant positive correlation for 10 out of 16 words. Table 2 presents the correlation coefficients and p-values obtained for each word.

	ρ	p
federal	0.131	0.001
spine	0.195	0.032
optical	0.227	0.003
compact	0.229	0.002
signal	0.233	0.008
leaf	0.252	0.001
net	0.361	0.001
coach	0.433	0.007
sphere	0.446	0.002
mirror	0.454	0.027
card	0.358	0.055
virus	0.271	0.159
disk	0.183	0.211
brick	0.203	0.263
virtual	-0.085	0.561
energy	0.002	0.990

Table 2: Correlation results per word.

Overview

Each question includes two sentences. Both sentences contain a **target word** between double brackets, as in: [[target]]. Your task is to rank the similarity of the two usages of the target word according to the following scale:

1. unrelated
2. distantly related
3. closely related
4. identical

IMPORTANT: your task is to evaluate the similarity of the two usages of the same word, **not the similarity of the two sentences** in general.

If you are unable to choose a label because you do not understand the sentences, select the option "cannot decide". Please try to use this option as little as possible!

An example

You will see two sentences. Both contain the target word marked by double brackets; in this example it's the word [[current]].

Read the sentences carefully:

- in any case , it 's not a question of electrocution . we can arrange a relay which will break the [[current]] at the instant of application of weight . if the robot should place his weight on it , he wo n't see .
- already , while it was still a blueprint , they were proud of their idea , of its simple clean lines and undeniable originality -- it owed nothing in its conception to any of the [[current]] models of revolutionary strategy . the japanese red army comrades , whadi haddad and his pflp contingent , even the matchless " carlos " could only admire .

And then select how similar the two usages of the word [[current]] are:

1. unrelated
2. distantly related
3. closely related
4. identical
5. (cannot decide)

You can choose only **one** label. Please try to use the option "cannot decide" as little as possible.

Why do texts look weird?

The sentences you'll read don't look like they were taken from a book. This is because they have gone through some text processing. You should not be concerned nor influenced in your decisions by the fact that:

- all words are lowercase (written in small letters), even proper names or the pronoun "I"
- whitespaces may appear where you don't expect them (e.g. before a comma) and may sometimes not appear where you'd expect them (e.g. between words)
- strange characters and words occasionally appear
- some words are misspelled
- a few words are missing
- the target word may appear multiple times (*but your judgement should be about the occurrence signalled by the*[[ ]] *marker*)

Please simply **ignore** these aspects while labelling!

Figure 4: Annotation instructions (part 1).

What do the similarity labels mean? More examples

Let's now look at examples for all four labels. Remember that you are evaluating the similarity of two word usages—not the overall similarity of the two sentences!

1. How similar are these two usages of [[current]]?

- prices of the leading issues . considering past earnings records , are apparently on a conservative basis measured by [[current]] market valuations in other groups . on the other hand there is no particular speculative incentive for operations in this group , with all signs pointing to a lower volume of sales in the last half of the year .
- one of the weirdest was the disappearance of anchovies off the coast of peru . why this happened is still unclear . one theory is that the cause was the 1972 - 73 invasion of a warm-water [[current]] called el nino , which upset the ecology of the coldwater humboldt current , drastically reducing the supply of 119 economics in plain english plankton and other nutrients on which anchovies (as well as whales) feed .

UNRELATED.
In the first sentence, "current" means being most recent or occurring at the present time. In the second sentence, "current" refers to a flow of water within a lake or an ocean. These two meanings have no properties in common; it is not possible to explain one usage in terms of the other.

2. How similar are these two usages of [[current]]?

- it is quite possible to arrive at the right conclusions for the wrong reasons , just as it is possible to ignore history but not to repeat it . thus , the summers book represents an important [[current]] of thought in the u.s. military , which rightly argues that the vietnam defeat was not the fault of the military ; never again should young americans be sent into battle without public backing and a clear definition of the goals of the military engagement .
- get busy with that l-tube ! If you do n't have it apart , cleaned , and together again before the day is out , i 'll coagulate your brains with alternating [[current]] . " not a robot moved !

DISTANTLY RELATED.
In the first sentence, "current" refers to a feeling or idea that exists within a group of people. In the second sentence, "current" refers to a flow of electricity. Although these two meanings of "current" are different, they are related as they do share some properties: for example, both currents of thought and electronic current can *flow*, and they are both often the result of an interplay of forces. This is why connecting the two meanings in a sentence results in a perhaps sophisticated but understandable statement: (A) "a new current of thought is flowing through the circuits of parliament".

3. How similar are these two usages of [[current]]?

- one of the weirdest was the disappearance of anchovies off the coast of peru . why this happened is still unclear . one theory is that the cause was the 1972 - 73 invasion of a warm-water [[current]] called el nino , which upset the ecology of the coldwater humboldt current , drastically reducing the supply of 119 economics in plain english plankton and other nutrients on which anchovies (as well as whales) feed
- he said he had survived by managing in a stupor to drag himself into a windowless shed behind his house. col. michael wiener , an israeli army doctor , said many of the survivors in the valley below the lake , such as in souboum , may simply have been in an air [[current]] that did not have any poison , while someone standing only a few yards away may have been killed. dr. weiner , the head of a 17-member rescue unit that came to cameroon team to nkamba , about 100 miles northeast of bamenda .

CLOSELY RELATED.
In the first sentence, "current" refers to a flow of water within a lake or an ocean. In the second sentence, "current" refers to a steady flowing movement of air. These two meanings are closely related: both refer to a steady and continuous flowing movement of some physical element. As in the previously encountered example (A), we can construct a sentence that relates the two usages: (B) "I can't remember whether El Nino is the name of an ocean or a wind current". Note, however, that the meanings are still ultimately different: a flow of water is not a flow of air.

4. How similar are these two usages of [[current]]?

- dell 's shares , on the other hand , go for 26 times projected 2004 earnings-but its business is three times as profitable as apple 's . The company 's supporters say [[current]] profits matter little because jobs has proved time and time again that he can create new products and trailblaze markets . that may be so , but as transamerica portfolio manager chris bonavico , who does n't own apple 's stock , notes , " Apple will remain a company that is neat from a product and consumer standpoint but crap from an investor standpoint .
- prices of the leading issues . considering past earnings records , are apparently on a conservative basis measured by [[current]] market valuations in other groups . on the other hand there is no particular speculative incentive for operations in this group , with all signs pointing to a lower volume of sales in the last half of the year .

IDENTICAL.
In both sentences, "current" means being most recent, up-to-date, or occurring at the present time. The two meanings are identical because they share virtually all properties, as can be seen from the following example. In the sentence (C) "I can't remember whether Donald Trump is the current president or vice-president of the United States", the meanings of "current" are ultimately the same regardless of whether Trump is president or vice-president. (Note the difference with respect to the constructed sentence (B) above.)

Now you're ready to start!

Figure 5: Annotation instructions (part 2).