# Measure and Evaluation of Semantic Divergence across Two Languages

**Syrielle Montariol** *
INRIA Paris
syrielle.montariol@inria.fr

**Alexandre Allauzen**
ESPCI Paris
Dauphine University
alexandre.allauzen@espci.psl.eu

## Abstract

Languages are dynamic systems: word usage may change over time, reflecting various societal factors. However, all languages do not evolve identically: the impact of an event, the influence of a trend or thinking, can differ between communities. In this paper, we propose to track these divergences by comparing the evolution of a word and its translation across two languages. We investigate several methods of building time-varying and bilingual word embeddings, using contextualised and non-contextualised embeddings. We propose a set of scenarios to characterize semantic divergence across two languages, along with a setup to differentiate them in a bilingual corpus. We evaluate the different methods by generating a corpus of synthetic semantic change across two languages, English and French, before applying them to newspaper corpora to detect bilingual semantic divergence and provide qualitative insight for the task. We conclude that BERT embeddings coupled with a clustering step lead to the best performance on synthetic corpora; however, the performance of CBOW embeddings is very competitive and more adapted to an exploratory analysis on a large corpus.

## 1 Introduction

Languages evolve throughout time: for many words, their usages along with their frequent collocations and associations can change, revealing the evolution of the society (Aitchison, 2001). However, all languages do not evolve identically: the impact of an event, the influence of a trend or thinking, can differ between communities. Moreover, languages do not evolve independently; some words can be inherited and borrowed between languages. For example, *cognates* — words that have the same etymological origin and similar meaning in two languages — can sometimes diverge into false friends, due to particular features of one language and its associated culture and history.

A more specific example is the Russian word "*ukrop*", meaning "dill". It started to be used by Russian people as an ethnic slur—a pejorative term—to talk about Ukrainian soldiers at the beginning of the Russian-Ukrainian conflict (Stewart et al., 2017). Then, Ukrainian people started to use it to designate their own patriots, in a positive way. Analysing the evolution of this word can lead to a better understanding of the evolution of the conflict; on the contrary, without suitable tools and methods to detect the divergence in its usage and connotation between communities, one might draw spurious results when analysing texts of this period.

Diachronic semantic change detection is an emerging field in Natural Language Processing, building upon the growing number of digitised texts with temporal metadata publicly available in various languages. It opens new perspectives of improvement for downstream tasks (using time-aware word representation for tasks ranging from text classification to information retrieval in temporal corpora) or for socio-linguistic and historical linguistics analysis (Kutuzov et al., 2018).

The goal of this paper is to extend the analysis of lexical semantic change across two languages, aiming at estimating the degree of diachronic semantic divergence between a word and its translation across time in a bilingual corpus. We propose an experimental framework to learn word representations that are comparable across both time and languages, and to detect and classify semantic divergence in a bilingual setting. We compare: *(i) diachronic* word embeddings, which allow static embeddings such as CBOW (Mikolov et al., 2013) to drift through time, and *(ii) contextualised* embeddings, relying on a pre-trained multilingual lan-

* This work was carried out while the author was working at LISN-CNRS.



*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1247–1258
August 1–6, 2021. 

guage model (M-BERT, Devlin et al., 2019).[1] We also propose an anchored-alignment strategy to tackle the bilingual setting for non-contextual embeddings. Then, we suggest a metric to measure the divergence of word usage between two languages, the *bilingual divergence*. Given the lack of a bilingual dataset annotated with semantic divergence, we generate a corpus of synthetic semantic drift across two languages using EuroSense (Delli Bovi et al., 2017), a sense-disambiguated and aligned bilingual corpus. To do so, we define a set of monolingual and bilingual semantic change scenarios and evaluate our different approaches on them. Finally, we apply our systems to newspaper corpora in two languages, English and French, covering the same time period, from 1987 to 2006. We classify all words of a bilingual lexicon into the scenarios defined for the synthetic drift generation.

To sum up, we extend the most appropriate methods from the literature of diachronic semantic change to build a framework for the measure of semantic divergence across languages (Sections 3 and 4), for which we propose a definition of the task, a measure of semantic divergence (Section 5), and a process to evaluate the presented methods (Section 6).

## 2 Related Work

**Diachronic embedding models.** The first approaches to diachronic modeling were based on relative word frequencies and distributional similarities (Gulordava and Baroni, 2011). Following the generalisation of word embeddings, *diachronic* word embeddings models emerged (Tahmasebi et al., 2018). A first line of work, led by Kim et al. (2014), learns an embedding matrix on the first time slice of a temporal corpus, and *incrementally* fine-tune it at each time step. This method has the advantage of simplicity but face a greater sensitivity to noise (Shoemark et al., 2019; Kaiser et al., 2020). Another method, proposed by Hamilton et al. (2016) and Kulkarni et al. (2015), train word embeddings on each time slice *independently* and align the representation spaces to make the embeddings comparable. Finally, Rudolph and Blei (2018); Jawahar and Seddah (2019) and Bamler and Mandt (2017) define probabilistic models of word embeddings, able to capture the drifts by training embeddings *jointly* on all time slices.

These methods average all the senses of a word into a unique vector at each time step. Pre-trained language models such as BERT (Devlin et al., 2019) allow each occurrence of a word to have a contextualised vector representation. These models, pre-trained on large datasets, improved the state of the art on numerous NLP tasks. Similarly, contextualised embeddings can be applied to semantic change detection (Giulianelli et al., 2020; Montariol et al., 2021) using several aggregation techniques to measure the degree of semantic change of a word from all its contextualised representations over time. However, these methods are still outperformed by non-contextualised embeddings for this task (Schlechtweg et al., 2020).

**Semantic change across languages.** While this topic is actively researched in the linguistic and sociology research communities (Boberg, 2012), it is fairly new in the NLP literature. Many authors apply diachronic embeddings models to more than one language (Hamilton et al., 2016; Schlechtweg et al., 2020). However, prior work comparing the evolution of word usage across languages is very limited. Some work studies variations between languages or dialects, without looking into the temporal dimension (Hovy and Purschke, 2018; Beinborn and Choenni, 2020). Uban et al. (2019) compare present meanings of cognate words across 5 Romance languages to differentiate true cognates from false friends and measure the divergence between languages. In a temporal fashion, Martinc et al. (2020a) study the evolution of 4 word pairs in an English-Slovenian corpus of newspaper articles. Finally, Frossard et al. (2020) propose a list of cognates for analysing the similarities in the evolution of English and French, along with a preliminary analysis focusing on the differences in word frequency over time.

## 3 Diachronic Words Embeddings

Before presenting systems based on contextualised embeddings, we introduce two methods using non-contextualised ones, as they are known to perform best for the task of semantic change detection (Schlechtweg et al., 2020). We use the continuous bag of words (CBOW) architecture of Word2Vec (Mikolov et al., 2013); we apply two different training methods to train it in a diachronic way. Then, we describe an *anchored-alignment* method to obtain bilingual diachronic word embeddings.

[1] Code is available at https://github.com/smontariol/BilingualSemanticChange

### 3.1 Diachronic Training

In this section, we consider a monolingual corpus divided into $T$ time slices. We rely on a fine-tuning method rather than an alignment-based method, where a new model would be trained from scratch at each time step (Hamilton et al., 2016). Indeed, for our cross-lingual task an alignment is already needed to map the embedding spaces of the two languages together; it would not be desirable to multiply this type of transformation, as each alignment introduces uncertainty in the system.

To begin with, as advised by Rudolph and Blei (2018), we pre-train our CBOW models on a shuffled version of the full corpus for each language. We use two methods for diachronic training. The first on is *incremental* training (Kim et al., 2014): we incrementally fine-tune the model on each time slice by initialising the weights with those of the previous time slice. The second variant is *independent* training: the model is fine-tuned on each time slice independently by initialising it with the pre-trained embeddings. Compared to the incremental method, the latter does not take into account the chronology of the corpus and can lead to less directed drifts. However, the fact that the embeddings do not go through a large amount of successive training updates, contrarily to the incremental method, prevents the embeddings from undergoing too extreme drifts (Shoemark et al., 2019).

### 3.2 Bilingual Alignment

We now consider a bilingual corpus, and embeddings trained separately on each language. We want to align the representation spaces to make the embeddings comparable.

**Anchoring.** The supervision signal for the alignment is key to the performance of the overall system, even more than the model architecture itself (Ruder et al., 2019). *Anchoring* is a form of supervision commonly used in NLP to obtain cross-lingual word embeddings. The supervision comes from a bilingual dictionary, whose words – the *anchors* – are used as seeds during the alignment. It can be transparent words such as named entities, or an exhaustive bilingual dictionary with the full vocabulary. However, aligning the vectors of the whole vocabulary is not appropriate for semantic change detection, as it tends to lower the disparities between the different vector spaces (Tsakalidis et al., 2019). In our case, the alignment forces the embeddings of the word pairs from the supervision dictionary to be the same in the two languages. This might hide some behavior such as a high disparity at the beginning of the full period and a convergence of meanings over time. Consequently, we use a seed dictionary with only the words that we assume are stable during the period in both languages. A first set of "stable" words are stopwords (Azarbonyad et al., 2017; Martinc et al., 2020b); however, by definition they do not carry much meaning. Relying only on them for the supervision might result in a poor alignment. We complement the list of seed words with word pairs that have the same relative frequency in the corpora of each language; with this frequency being in the top 10% of the full corpus (Azarbonyad et al., 2017; Zhang et al., 2015). For all experiments in this paper, we use the bilingual dictionary from the MUSE tool[2] (Lample et al., 2018). It includes 5000 word pairs and handles word polysemy.

**Alignment.** First, we train monolingual CBOW embeddings on each language independently, without dividing the corpora into time slices. To prepare for the alignment, we apply mean-centering to the embeddings of each language, as Schlechtweg et al. (2019) showed the positive impact of this preprocessing step for vector space alignment. For the alignment, we use Orthogonal Procrustes (Schönemann, 1966). It consists in finding the mapping $W$ between two embedding spaces $E_1$ and $E_2$ which minimizes the sum of squared Euclidean distances between the image of the source embeddings space $E_1 * W$ and the target embedding space $E_2$ for the set of selected anchor words in both spaces. These aligned embedding vectors are used to initialise the diachronic embeddings, which can then be trained on all the time slices in both languages, incrementally or independently.

## 4 Contextualised Embeddings

To challenge the systems based on aligned CBOW embeddings, we use M-BERT, the multilingual version of BERT (Devlin et al., 2019). It is trained on Wikipedia content on 104 languages, without any additional multilingual mechanism nor language identifier.

Applying a pre-trained multilingual model on a bilingual temporal corpus enables immediate comparison without requiring any alignment. Each sequence is labelled with the time it was written and

[2] https://github.com/facebookresearch/MUSE

its language. We extract contextualised representations for each token of a sequence by summing the top four hidden layers of the pre-trained model. BERT representation relies on a system of wordpieces; if a word is divided into several wordpieces, we take the average of all the wordpiece embeddings as representation for the word. To sum up all the information about a word from the set of contextual embeddings of all its occurrences in a time slice, we experiment with two aggregation techniques: averaging and clustering.

**Averaging :** Proposed by Martinc et al. (2020a), this method averages all the token embeddings of a word for each time period and each language. We end up with a set of time-specific and language-specific vector representations of a word. They can be compared using the cosine distance (Shoemark et al., 2019).[3]

**Clustering:** This method, first used by Giulianelli et al. (2020), groups the set of token embeddings of a word into types of usages. We apply a clustering algorithm, k-means, to all the embeddings of a word and its translation, on all the time periods jointly. Then, we compute the normalised distributions of clusters, for each language and period. More precisely, for a given word, we extract the number of tokens in each cluster and for each pair (period, language); we normalise it by the total number of occurrences of the word in the corpus. We obtain the probability distributions of the usages of this word at each time slice and in both languages. These distributions can be compared between two periods or two languages using the Jensen-Shannon divergence (JSD, Lin, 2006).

## 5 Drift Measures

After applying the described systems to a bilingual corpus divided into $T$ time slices, for a given target word in a given language $l$, we obtain either a sequence of $T$ embeddings $\boldsymbol{u}_l^{(t)}$ in each language (for CBOW and m-BERT with averaging), or a vector of $T$ cluster distributions $\boldsymbol{c}_l^{(t)}$ (for m-BERT with clustering). We compute the distance between representations: the cosine distance between non-contextual embeddings and the JSD between clusters distributions.

$$d(t_1, t_2, l_1, l_2) = \begin{cases} \cos(\boldsymbol{u}_{l_1}^{(t_1)}, \boldsymbol{u}_{l_2}^{(t_2)}) & \text{(averaging} \\ & \text{or CBOW)} \\ \text{JSD}(\boldsymbol{c}_{l_1}^{(t_1)}, \boldsymbol{c}_{l_2}^{(t_2)}) & \text{(clustering)} \end{cases} \quad (1)$$

In a monolingual setting, we use two metrics commonly used to measure the drifts of a word in each language (Rodina et al., 2019): the *incremental* drift, from each time slice to the next one, and the *inceptive* drift, from the beginning of the period to each time slice. We obtain drift vectors in $\mathbb{R}^{T-1}$ for each word in each language, by computing $d(t_1, t_2, l, l)$.

In a bilingual setting, drift measures can be computed for each word pair (one word and its translation). First, we compute the distance inside each word pair at each time step. We call it the bilingual distance: $s_B^{(t)} = d(t, t, l_1, l_2)$ for $t = 1, 2, \ldots, T$. Second, the temporal drift of this distance is measured similarly to the monolingual drift, either *incrementally* or *inceptively*. The distance is the norm between the bilingual distance $s_B^{(t)}$ at two time steps, measuring the divergence of the usage of a word and its translation. We call it *bilingual divergence*. For example, the *incremental* bilingual divergence is computed as follows:

$$D_B^{\text{incr}} = \begin{pmatrix} |s_B^{(0)} - s_B^{(1)}| \\ |s_B^{(1)} - s_B^{(2)}| \\ \vdots \\ |s_B^{(T-1)} - s_B^{(T)}| \end{pmatrix} \quad (2)$$

Various information can be extracted from the vector of bilingual divergence of a word $D_B$: the trend (*no trend* i.e. stable distance between a word and its translation, *decreasing* i.e. convergence, or *increasing* i.e. divergence), the degree of divergence (e.g. by summing all its elements), and the speed of divergence (by estimating the slope).

## 6 Synthetic Drift Generation

The study of semantic change faces the issue of evaluation, as few labeled corpora exist for this task. Recent initiatives from the NLP community start to produce more annotated data (Schlechtweg et al., 2020); however, no corpus is available for bilingual analysis. Consequently, we generate a corpus of bilingual synthetic semantic change, following common practice in the literature of monolingual semantic change detection (Shoemark et al.,

[3] We define the cosine distance as (1 - cosine similarity).

2019; Schlechtweg and Schulte im Walde, 2020). It allows us to control exactly the shape and degree of semantic change in the corpus and thus gain a deeper understanding of the impact of each modeling choice.

To create synthetic semantic change, common practice involve to merge two words that do not share a common sense, creating a pseudo-word; then, generate synthetic change by controlling the proportion of sentences using each of the two original words in the successive time slices of a temporal corpus (Rosenfeld and Erk, 2018; Shoemark et al., 2019). However, as advised by Schlechtweg and Schulte im Walde (2020), it is preferable to use the natural polysemy of words for the synthetic drift to be as close as possible to reality: instead of controlling the proportion of sentences containing two unrelated words merged as a pseudo-word, we use sentences containing several senses of a unique word. To this end, we need a bilingual sense-annotated corpus with consistent annotations between languages (Pasini and Camacho-Collados, 2020). The EuroSense corpus[4] (Delli Bovi et al., 2017) is derived from the Europarl corpus, a large public corpus of proceedings of the European Parliament. It has a full and a refined version. We use the latter to build our synthetic corpus; it is half the size of the first one but more reliable. The framework BabelNet (Navigli and Ponzetto, 2012) is used for annotation. EuroSense contains parallel text in 21 European languages. We focus on the two languages with the highest amount of annotations in the refined corpus: English and French. An example of aligned sentences in these languages can be found in Table 1.

| | English | French |
|---|---|---|
| Sentence | The best tools for this are liberalisation and freer competition , which causes train companies to take a greater interest in the wishes of customers . | *Les meilleurs moyens d'y parvenir sont la libéralisation et une concurrence plus libre , qui incite les compagnies ferroviaires à se soucier davantage des souhaits de leurs clients .* |
| Lemma | customer | *client* |
| Sense | bn:00019763n | bn:00019763n |

Table 1: Example of aligned sentences in English and French in the EuroSense corpus, with annotated anchor and corresponding sense in the BabelNet framework.

### 6.1 Semantic change Scenarios

In order to generate and capture variations of distributions of word senses through time and across two languages, we define several scenarios of word usage variations. First, we choose two monolingual scenarios of semantic change (labeled "M") and generate them using sentences extracted from the EuroSense corpus. Assuming we have a target word with at least two senses, the scenarios are:

- $M0$: all senses are fully stable.
- $M1$: one sense gradually appears / disappears, the others stay stable.

Second, we define scenarios of semantic divergence (bilingual scenarios, labeled "B") derived from the monolingual scenarios. Assuming we have a target words $w_1$ and its translation $w_2$ with at least two senses in common:

- $B0$: $w_1$ and $w_2$ are $M0$.
- $B1$: $w_1$ is $M0$, $w_2$ is $M1$.
- $B2$: $w_1$ and $w_2$ are the same $M1$ (they gain/lose the same sense, drifting in the same direction).
- $B3$: $w_1$ and $w_2$ are different $M1$ (one gains/loses one sense, the other gains/loses another sense: they diverge).

These 4 scenarios can be linked with distinct phenomena. Examples of words for each of them, extracted from a bilingual English-French corpus of newspaper articles spanning 20 years, can be found in Table 3. First, scenario $B0$ deals with words which have a stable meaning and an equivalent word with equally stable meaning in the other language (e.g. *dinosaurs*). Scenario $B1$ can be caused by a word being borrowed from one language to another: a *loanword*. After the borrowing, its usage can evolve, for example due to sociocultural specificity impacting the second language, while it stays stable in the source language. Similarly, an example of $B3$ scenario are cognate words whose usage evolve in their respective languages, diverging into false friends. For example, the English noun *affair* has common etymology with old French and used to mean "*what one has to do, ordinary business*". Its usage evolved across time,

[4] http://lcl.uniroma1.it/eurosense/

gaining in English the new sense of "*a love relationship, usually secret*" while it often refers in French to "*a business case*". The word *ukrop* presented in the introduction is also an example of $B3$ scenario. Finally, scenario $B2$ deals with words that go through the same semantic change as their equivalent in another language. Among other phenomenon, a common cause is when a language evolution is triggered by a cultural or technological change that is common to the societies speaking the two languages. For example, the sense of the word *confinement* related to pandemic became the majority meaning in many languages worldwide following the COVID-19 pandemic.

### 6.2 Building the Synthetic Corpus

**Step 1: selection of target lemma pairs.**

For all the sense-annotated lemmas in English and French in EuroSense, we extract their sets of senses. We only keep the senses with more than 200 occurrences per language. We associate English and French lemmas together if they have at least two senses in common, creating a bilingual dictionary. From these lemma pairs, we extract the set of sentences annotated with one of the senses in common to build the pool of sentences for the next step. In total, we have 115 English-French lemma pairs, of which 66 have 2 senses (low polysemy) and 49 have between 3 and 5 senses. For example, a low-polysemy lemma pair is (project, *projet*) and a high-polysemy one is (measure, *mesure*).

**Step 2: creation of sense distributions.**

For each monolingual scenario, we create probability distributions of senses at each time slice. We denote by $p(S \mid \mathcal{T}, W, L)$ the probability that the lemma $W$ conveys the sense $S$ at time $\mathcal{T}$ in language $L$. We generate $T = 10$ time slices and apply each scenario to all the target lemmas pairs. Since our variables are discrete, for a given lemma $w$ in language $l$, the probability distribution of a set of 2 senses $\{s_1, s_2\}$ over time can be characterised by a $2 \times T$ stochastic matrix, where the lines sum to 1:

$$\begin{pmatrix} p(s_1 \mid \mathcal{T}=1,w,l) & p(s_2 \mid \mathcal{T}=1,w,l) \\ p(s_1 \mid \mathcal{T}=2,w,l) & p(s_2 \mid \mathcal{T}=2,w,l) \\ \dots & \dots \\ p(s_1 \mid \mathcal{T}=T,w,l) & p(s_2 \mid \mathcal{T}=T,w,l) \end{pmatrix}.$$

For a given target lemma, for the $M0$ scenario, we randomly draw an initial distribution over the set of senses and repeat it at each time slice: $p(S \mid \mathcal{T} = t, w, l) = p(S \mid \mathcal{T} = 1, w, l)$ for $t = 2, 3, \dots, T$. For the $M1$ scenario, we gradually increase or decrease the probability of appearance through time of one of the senses, either linearly or logarithmically, following Shoemark et al. (2019). The other senses have a stable distribution across time.

**Step 3: creation of the synthetic corpus.**

For each monolingual scenario, we build the synthetic corpus time slice after time slice, using the set of target lemmas, the pool of sense-annotated sentences and the generated distributions of senses.

For each target lemma, at each time step $t$, we sample 200 sentences for each of its senses. Then, we add each sampled sentence to time step $t$ with the probability specified in the corresponding distribution of senses of the scenario. To avoid the synthetic sense distribution for a target lemma to be disturbed by noise from its appearance as a context word in other sentences, when adding a sentence to the synthetic corpus, we attach the suffix "_l" to its target lemma. Note that the 200 sentences sampled for each sense of a lemma can appear only once in each time slice, but can appear in other time slices of the corpus.

All the bilingual scenarios are built from the monolingual ones. Generating them reduces to using the right monolingual scenarios for each word and its translation. For example in the $B3$ scenario, we generate a corpus using the $M1$ scenario for both the target lemma and its translation, but select a different sense to appear or disappear in order to induce a divergence. The synthetic corpora, for each scenario and each language, have around 7.5M words distributed into the 10 time slices.

### 6.3 Evaluation Method

To sum up, at each time $t$, a word $w$ in a language $l$ is characterised by its sense distribution in the synthetic corpus $p(S \mid t, w, l)$. This information is similar to the cluster distributions extracted when applying clustering to contextualised embeddings; we can compute the drift measures defined in Section 5, using the JSD to compare the sense distributions. The drifts obtained from these measures can then be used as gold standard for the evaluation of our systems.

For each system described in sections 3 and 4 and for each target lemma pair, we output the vectors of *monolingual drift* computed on the monolingual scenario synthetic corpora and the vectors of *bilingual divergences* computed for the bilin-

| Model | Diachrony | Stable $M0$ | Drift $M1$ | Both stable $B0$ | Stable&drift $B1$ | Same drift $B2$ | Diverge $B3$ |
|---|---|---|---|---|---|---|---|
| CBOW | incremental | 0.65 - 0.16 | 0.54 - 0.96 | 0.87 - 0.82 | 0.66 - 0.46 | 0.76 - 0.68 | 0.63 - 0.47 |
| | independent | 0.84 - 0.83 | 0.63 - 0.86 | 0.83 - 0.89 | 0.70 - 0.45 | 0.80 - 0.66 | 0.67 - 0.50 |
| BERT | averaging | 0.86 - 0.87 | 0.34 - 0.55 | 0.84 - 0.90 | 0.79 - 0.4 | 0.71 - 0.69 | 0.63 - 0.47 |
| | k-means 5 | 0.85 - 0.86 | 0.61 - 0.19 | 0.86 - 0.97 | 0.78 - 0.41 | 0.77 - 0.91 | 0.66 - 0.40 |

Table 2: Accuracy measure of each system for the different semantic change scenarios. The numbers on the left are *incremental drift* while the ones on the right are *inceptive drift*.

gual scenarios (see Section 5). We wish to evaluate whether these series have the same trend as the gold standard. For this, we use the Mann-Kendall (MK) Trend Test (Mann, 1945; Kendall, 1975), a non-parametric statistical test used to detect trends of variables. It is particularly suited to monotonic trends, which is how we designed the semantic drift in our data.

The null hypothesis of the test is the absence of monotonic trend. The Mann-Kendall test statistic $Z_{MK}$ relies on comparing every value in the time series with all the values preceding it. The sign of the statistic test indicates the trend of the data, given a confidence level of 0.05: no monotonic trend (the null hypothesis), increasing trend ($Z_{MK} > 0$), or decreasing trend ($Z_{MK} < 0$). For a given target lemma, if the direction of the detected trend in our data is the same as the one from the gold standard drift, we consider that the semantic change has been correctly identified. We compute the accuracy as the proportion of correctly identified trends in the full list of target lemmas.

## 7 Experiments on Synthetic Data

We compare the accuracy of our systems on the synthetic corpora generated in the previous section.

### 7.1 Experimental Setup

**CBOW processing.** As we rely on stopwords (on top of frequent words) for the alignment, we do not discard them during preprocessing. The context size is set to 5 words, and the dimension of word embeddings to 50. Preliminary experiments with larger embedding dimensions exhibited no significant improvement. We posit this is due to the small size of the dataset. Moreover, the accuracy of incremental fine-tuning of CBOW embeddings for semantic change detection is very sensitive to dimensionality (Kaiser et al., 2020); the optimal embedding dimension is usually quite low, with a clear drop in performance with high embeddings dimensions. We train all models using 10 epochs. For each language, a static model is first trained on the set of all sentences containing the target lemmas. Then, we proceed with *incremental* an *independent* training.

**BERT processing.** We use the pre-trained *bert-base-multilingual-uncased* model from the `transformers` library. We extract the contextualised embeddings from the corpus and apply the two aggregation methods, *averaging* and *clustering*. We choose $k = 5$ clusters for k-means, as it is the maximum number of senses that can be found in our list of target lemmas. Experiments with higher values of $k$ did not improve the accuracy. We remove the "_l" suffix of the target lemmas before extracting their embeddings.

### 7.2 Results on synthetic data

Table 2 summarises the accuracies measured using the Mann-Kendall trend test (Hussain and Mahmud, 2019) on the 115 lemma pairs. It compares the trend of the drift of all systems with the gold standard trend, for each scenario. We have three scenarios with stable monolingual drift or stable bilingual divergence ($M0$ and $B0$, with all the senses being stable; and $B2$, where a word and its translation drift in the same direction) and three drifting scenarios ($M1$ and $B1$, where one sense drifts; and $B3$, where a word and its translation drift in different directions). The results show that stable scenarios are generally easier to detect accurately compared to the changing ones, especially in the monolingual analysis.

The best results are obtained with BERT using k-means clustering. This system focuses on the variation of proportion of the different usages, instead of the evolution of the average word representation; it provides a better focus on the meaningful changes in word usage. In the case of CBOW, independent training leads to better performances than incremental training. This is in line with the find-

ings of Shoemark et al. (2019): the large amount of training updates, especially in such a small corpus, is harmful for the quality of the representation.

Overall, the *inceptive* drift measure leads to better accuracy for stable scenarios, while the *incremental* drift is more suited to scenarios where the sense distributions change across time. Thus, we advise towards always computing both measures for diachronic studies.

## 8 Experiment on Newspaper Corpora

We analyse the semantic divergence of word-translation pairs in a bilingual corpus of news articles. Our goal is to classify all words of a bilingual lexicon into the semantic divergence scenarios defined in Section 6.1.

### 8.1 Data Description & Experimental Setup

The *New York Times Annotated Corpus* (Sandhaus, 2008) gathers around 1 855 000 articles from January 1987 to June 2007. We scrape *Le Monde*, one of the most read daily newspapers in France, on the same time period. We divide both corpora into $T = 20$ yearly time steps, as a trade-off between getting precise information on semantic drift thanks to a low granularity and reducing noise that appears due to a too low granularity. Finally, we select a vocabulary containing the $V = 40\,000$ most frequent words for each corpora. The average number of words is around 3.5 M for one time step in the French corpus and 9 M in the English one. First, a bilingual lexicon is built using the intersection of the MUSE bilingual dictionary with the French and English vocabularies from our corpora. We manually update the bilingual lexicon with domain-specific vocabulary such as named entities, in order to improve the coverage on the corpora. The final bilingual dictionary has 27 351 words.

To obtain bilingual diachronic embeddings, we use CBOW with incremental training. Indeed, even though BERT with k-means clustering lead to better results overall on synthetic corpora, the extraction of each token embedding and the clustering step are computationally heavy. Moreover, in a large corpus such as ours, saving in memory as many embedding vectors as occurrences of words from the bilingual lexicon is not feasible. Thus, the clustering method is more suited for a fine-grained analysis of the divergence of senses of a limited set of target words, rather than an exploratory analysis on the full vocabulary.

The experimental setup is the same as the one used on the synthetic corpus; the volume of data being much higher in the newspaper corpus, we increase the capacity of our model by setting the dimension of CBOW embeddings to 100, in order to retain more information. We pre-train CBOW models on the English and French corpora and normalise the embeddings to prepare for the alignment. The French corpus being the smallest, its embeddings are mapped to the English embedding space. Then, we incrementally update the aligned embeddings on both corpora. For each word of the bilingual vocabulary, we compute its monolingual drift and its bilingual divergence, following the methodology applied on the synthetic corpora. It allows us to identify the words belonging to each of the bilingual scenarios.

### 8.2 Results on Bilingual Divergence

On top of classifying all words into the different bilingual divergence scenarios, we quantify the degree of divergence by summing up the elements of the vectors of *inceptive drift* and of *inceptive bilingual divergence* respectively. The proportion of each scenario as well as examples selected among the words with the most extreme drifts are in Table 3. For example, words belonging to scenario $B3$ have the highest monolingual drifts in both the English and French corpora, while their bilingual divergence is among the lowest.

Words that are stable in both languages ($B0$) are mostly daily life words (e.g. *mayonnaise*). Words that drift in the same direction in both languages ($B2$) are concepts related to technology and society that are common to the English and French culture (e.g. *renewable*); while the words that diverge between the two languages ($B1$-fr (English stable, French drifting), $B1$-en and $B3$) belong to more culture-specific concepts (e.g. *francs*) or controversial topics (e.g. *terrorist*). For example, *francs* drifts in French, while it is stable in English. This is probably due to the change of currency in France in 2002 that had much lower media coverage in the US. Similarly, *terrorist* drifts in both languages but in different directions. Indeed, the two countries went through many terrorist attacks during the period under study, but from very different groups, leading to different contexts for this word.

Overall, the exploratory results on the bilingual newspaper corpora offer interesting insights on perspectives for many applications; both for long-term

| B0: both stable 58.2% | B1-fr: stable&drift 15.5% | B1-en: drift&stable 16.2% | B2: same drift 4.9% | B3: different drifts 5.2% |
|---|---|---|---|---|
| dinosaurs | reforms | bush | genomics | steroid |
| pottery | delinquency | horrific | renewable | rockets |
| anniversaries | francs | maid | condom | gay |
| mayonnaise | feminine | hostages | cinemas | katrina |
| joke | provincial | dealers | robotic | terrorist |

Table 3: Proportion and example words for the different categories of bilingual divergence.

semantic change, studying the joint evolution of cognate words and borrowings; and for short-term change in word usage, for example when studying the disparity in the media resonance of an event in different countries.

## 9 Discussion

In this paper, we define an experimental framework to measure and classify the semantic divergence of a word and its translation in a bilingual corpus. We compare different kinds of word embeddings on various bilingual divergence scenarios generated in a synthetic corpus. We apply our conclusions to a bilingual newspaper corpus to identify words undergoing different types of semantic divergence. BERT embeddings coupled with a clustering step lead to the best performance on synthetic corpora. The performance of CBOW embeddings is nevertheless very competitive, and more adapted to an exploratory analysis on a large corpus.

There is a large margin for future work; be it in terms of quality of diachronic bilingual representation, metric to measure semantic divergence, and evaluation method. Our evaluation focuses on the trend of the drift, but its degree and its speed can also be quantified and analysed. In addition, the underlying bilingual representation learning approach is key for the detection of drifts. The transformations applied to create a cross-lingual word embedding space might result in information loss or generation of spurious drifts in the embeddings. To compare word embeddings with the purpose of detecting semantic divergence, the anchored alignment method presented here is not the only option; promising candidates are Temporal Referencing (Schlechtweg et al., 2019) and the Global Anchor method (Yin et al., 2018).

A limitation of our work is the use of an injection to define word pairs. In his general linguistics course, De Saussure (1916) states that there is no bijective relationship between words in different languages. The different meanings and uses of a word in a language cannot have a perfectly identical equivalent in another language. Moreover, as noted by Frossard et al. (2020), a word can have synonyms in one language while the word bearing the same meaning in another language has none; in that case, the usage of the word in the first language is divided into all its synonyms.

Another limitation is evaluation with synthetic data. This method is common in monolingual semantic change analysis, but there is no guarantee that the generated phenomenon is similar to real-world data. For example, a degree of freedom is the shape of the synthetic drifts generated. In this paper, we used logarithmic and linear shapes; but some literature hint that a logistic shape is also a good match for semantic drift (Bailey, 1973; Blythe and Croft, 2012). Furthermore, in real data the granularity (the size of the periods used to divide the corpus) might have an important impact on the shape of the semantic evolution.

Finally, as we build all bilingual scenarios from combinations of two monolingual scenarios, the flaws of the monolingual scenarios are inherited by the bilingual scenarios. It can potentially multiply the noise by propagation of uncertainty. We wished to overcome the limitations of synthetic evaluation with the application on real corpora, but more thorough interpretation would be necessary for a solid qualitative evaluation. To perform quantitative evaluation on real data, an annotated dataset similar to the ones for monolingual semantic change (e.g. Schlechtweg et al., 2020) would be necessary. However, the annotation task would be even more complex than for monolingual data. An easier entrance point towards annotating data for this task could be loanwords and cognate words. Overall, this is a challenging task and we hope to attract more people to work on it in the future.

## References


Jean Aitchison. 2001. Language change: Progress or decay? In *Cambridge Approaches to Linguistics*. Cambridge University Press.

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. 2017. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 1509–1518. Association for Computing Machinery.

Charles-James N Bailey. 1973. Variation and linguistic theory.

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 380–389, International Convention Centre, Sydney, Australia. PMLR.

Lisa Beinborn and Rochelle Choenni. 2020. Semantic drift in multilingual representations. *Computational Linguistics*, 46(3):571–603.

Richard A Blythe and William Croft. 2012. S-curves and the mechanisms of propagation in language change. *Language*, pages 269–304.

Charles Boberg. 2012. English as a minority language in quebec. *World Englishes*, 31.

Ferdinand De Saussure. 1916. Cours de linguistique générale.

Claudio Delli Bovi, Jose Camacho-Collados, Alessandro Raganato, and Roberto Navigli. 2017. EuroSense: Automatic harvesting of multilingual sense annotations from parallel text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 594–600, Vancouver, Canada. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Esteban Frossard, Mickael Coustaty, Antoine Doucet, Adam Jatowt, and Simon Hengchen. 2020. Dataset for temporal analysis of English-French cognates. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 855–859, Marseille, France. European Language Resources Association.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

Md. Hussain and Ishtiak Mahmud. 2019. pymannkendall: a Python package for non parametric mann kendall family of trend tests. *Journal of Open Source Software*, 4(39):1556.

Ganesh Jawahar and Djamé Seddah. 2019. Contextualized diachronic word representations. pages 35–47.

Jens Kaiser, Dominik Schlechtweg, Sean Papay, and Sabine Schulte im Walde. 2020. IMS at SemEval-2020 task 1: How low can you go? dimensionality in lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 81–89, Barcelona (online). International Committee for Computational Linguistics.

Maurice G. Kendall. 1975. Rank correlation measures.

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. Association for Computational Linguistics.

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*, WWW '15, pages 625–635, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

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. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jianhua Lin. 2006. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theor.*, 37(1):145–151.

Henry B. Mann. 1945. Nonparametric tests against trend. *Econometrica: Journal of the econometric society*, pages 245–259.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020a. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020b. Discovery team at SemEval-2020 task 1: Context-sensitive embeddings not always better than static for semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 67–73, Barcelona (online). International Committee for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.

Roberto Navigli and Simone Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Tommaso Pasini and Jose Camacho-Collados. 2020. A short survey on sense-annotated corpora. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5759–5765, Marseille, France. European Language Resources Association.

Julia Rodina, Daria Bakshandaeva, Vadim Fomin, Andrey Kutuzov, Samia Touileb, and Erik Velldal. 2019. Measuring diachronic evolution of evaluative adjectives with word embeddings: the case for English, Norwegian, and Russian. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 202–209, Florence, Italy. Association for Computational Linguistics.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1003–1011, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Evan Sandhaus. 2008. The New York Times annotated corpus. In *Philadelphia : Linguistic Data Consortium*. Vol. 6, No. 12.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

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

Dominik Schlechtweg and Sabine Schulte im Walde. 2020. Simulating lexical semantic change from sense-annotated data. *CoRR*, abs/2001.03216.

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

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of EMNLP-IJCNLP 2019*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

Ian Stewart, Dustin Arendt, Eric Bell, and Svitlana Volkova. 2017. Measuring, predicting and visualizing short-term change in word representation and usage in vkontakte social network. In *ICWSM*.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to diachronic conceptual change. *ArXiv*, abs/1811.06278.

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

Ana Uban, Alina Maria Ciobanu, and Liviu P. Dinu. 2019. Studying laws of semantic divergence across languages using cognate sets. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 161–166, Florence, Italy. Association for Computational Linguistics.

Zi Yin, Vin Sachidananda, and Balaji Prabhakar. 2018. The global anchor method for quantifying linguistic shifts and domain adaptation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9434–9445, Red Hook, NY, USA. Curran Associates Inc.

Yating Zhang, Adam Jatowt, Sourav Bhowmick, and Katsumi Tanaka. 2015. Omnia mutantur, nihil interit: Connecting past with present by finding corresponding terms across time. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 645–655, Beijing, China. Association for Computational Linguistics.