Transparent Semantic Change Detection with Dependency-Based Profiles

Bach Phan-Tat[1], Kris Heylen[1,2], Dirk Geeraerts[1], Stefano De Pascale[1,3], Dirk Speelman[1]

[1]Department of Linguistics, KU Leuven
[2]Instituut voor de Nederlandse Taal
[3]Vrije Universiteit Brussel

Correspondence: ttbach.phan@kuleuven.be

Abstract

Most modern computational approaches to lexical semantic change detection (LSC) rely on embedding-based distributional word representations with neural networks. Despite the strong performance on LSC benchmarks, they are often opaque. We investigate an alternative method which relies purely on dependency co-occurrence patterns of words. We demonstrate that it is effective for semantic change detection and even outperforms a number of distributional semantic models. We provide an in-depth quantitative and qualitative analysis of the predictions, showing that they are plausible and interpretable.

1 Introduction

Lexical semantic change detection is a well-established field in Natural Language Processing (NLP), with several shared tasks for different languages, amongst which the most popular one is SemEval 2020 Task 1 (Schlechtweg et al., 2020). Most of the models use either static word embeddings like word2vec (Mikolov et al., 2013) or ELMo (Peters et al., 2018) or contextualised ones such as BERT (Devlin et al., 2019) or S-BERT (Reimers and Gurevych, 2019) which is the current state-of-the-art (Cassotti et al., 2023). Despite the differences in their architectures, they all build upon the distributional semantics hypothesis (Firth, 1957) and all indirectly represent meanings as numerical vectors, abstracting away from the level of individual co-occurring items. Yet, linguists have long observed that we can study semantics and word-usage by directly looking at the untransformed co-occurrence contexts of the target (McEnery et al., 2019; Stubbs, 1995), which suggests that we could study semantic change by looking at shifts in the contexts of the target. However, most co-occurrence methods, including a recent one by McEnery et al. (2019), rely on surface co-occurrences (i.e., bag-of-words approaches to
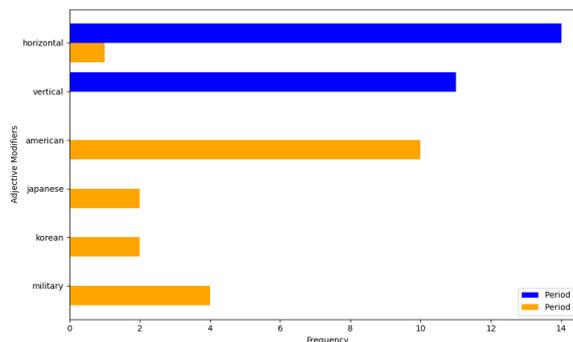

Figure 1: Changes in the frequencies of 6 adjectival modifiers of the English noun *plane* between the 2 sub-corpora for English of the SemEval 2020 shared task 1 (Schlechtweg et al., 2020). *Plane* is annotated as semantically changed in the dataset

co-occurrence). Surface co-occurrence often suffers from accidental and/or indirect co-occurrences and the arbitrary choice of the span size (Evert, 2008). Syntactic (dependency) co-occurrence (Evert, 2008; Garcia and García Salido, 2019; Seretan, 2011), with its different dependency relations, is more helpful to separate the signals. Consider the English noun *plane*: in the 20th century, the AIR-CRAFT meaning emerged and became more dominant with the rise of such adjectival modifiers as *american*, *korean*, *japanese*, *military* whereas the DIMENSIONAL meaning became less dominant with the decrease of such adjectival modifiers as *horizontal*, *vertical*, as shown in Figure 1.

Using distributional shifts of the dependency co-occurrences to detect lexical semantic change is the focus of this paper. We investigate to what extent we could use explicit linguistic information to detect semantic change. Our main hypothesis is that the degree of semantic change of a word is reflected in changes in its dependency co-occurrences. Note that we are not trying to set a new state-of-the-art but to show the usefulness of explicit linguistic information and more simple, transparent techniques.

We investigate this question experimentally[1] using the dataset for English, German, Swedish, Latin from SemEval 2020 Task 1. Our main findings are the following:

1. Tracing the changes in the distribution of dependency co-occurrences outperforms a number of distributional models.

2. As our method is transparent by design, we can easily understand how a word has changed (as in the *plane* example above). In contrast, dense embedding models often require post-hoc interpretability methods or additional analyses (e.g., Integrated Gradients (Sundararajan et al., 2017), LIME (Ribeiro et al., 2016) or SHAP (Lundberg and Lee, 2017), see surveys in Madsen et al., 2022; Arrieta et al., 2019) to obtain comparable, feature-level explanations. This makes our approach particularly convenient for linguistic and humanities studies where the explanation needs to be directly inspectable.

3. Our approach is compatible with linguistic theories of lexical semantics and semantic change (e.g., Geeraerts et al., 2023; Geeraerts, 1997; Hanks, 2013; Dalpanagioti, 2019), allowing researchers to bridge the gap between NLP and diachronic linguistics.

2 Related work

A long tradition in distributional semantic models treat lexical meaning as a function of usage: words are characterised by the contexts they appear in, and semantic change is operationalised as a shift in these contextual distributions over time. Within this tradition, work roughly splits into two dominant methodological styles.

The first tracks change directly in interpretable co-occurrence evidence (e.g., shifts in salient collocates or in the distribution of context features). This line has strong roots in mid-20th-century structuralism, combining Harris's distributional analysis (Harris, 1954) with Firth's contextual theory of meaning (Firth, 1957). With the rise of corpus linguistics and concordancing, this idea became a practical methodology. Corpus linguists compared a target word's prominent collocates and recurring patterns across corpora and later across time

slices, often supported by association measures for identifying characteristic collocations (Church and Hanks, 1990; Sinclair, 1991). This collocational tradition also developed notions such as semantic preference/prosody, which are systematic attitudinal or functional tendencies inferred from a word's typical collocational environments (Partington, 1998; Stubbs, 1995). This naturally lends itself to diachronic analysis when those environments shift. This is done by building time-sliced collocation profiles and examining how association-ranked collocates emerge, fade, or are replaced across periods, keeping the evidence directly inspectable. Case studies have used such collocation turnover to document long-term changes in conventionalised combinations (e.g., Alba-Salas, 2007; Pettersson-Traba, 2021). Some of the more recent semi-automatic methods were developed by Garcia and García Salido (2019) and McEnery et al. (2019). Different tools have also made this style scalable, enabling extraction and comparison of diachronic collocation phenomena in large corpora such as AntConc (Anthony, 2014), #Lancsbox (Brezina et al., 2021), SketchEngine (Kilgarriff et al., 2014), DiaCollo (Jurish, 2015; Jurish and Nieländer, 2020).

The second maps co-occurrence evidence into vector-space representations (count-based or neural embeddings). The first models used an untransformed co-occurrence matrix (later changed to other measures such as Mutual Information (MI) or Log Likelihood Ratio (LLR)), with each row corresponding to a word and the columns words in the vocabulary (Bullinaria and Levy, 2007; Gulordava and Baroni, 2011; Schütze, 1998). The result is that each word will be represented by a sparse count vector. Predictive neural embedding architectures like word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), FastText (Bojanowski et al., 2017) could be said to capture similar information as the count-based approach but are more powerful at abstracting information from the corpus with denser vectors (Baroni et al., 2014). Contextualised models such as BERT (Devlin et al., 2019) are even more powerful in the sense that they can distinguish between individual instances of word usages instead of representing one word as one vector (Tahmasebi and Dubossarsky, 2023). Several attempts have been made to integrate explicit linguistic information into neural network models, such as Dependency Skip-gram (Levy and Gold-

---

[1]All the codes used in this paper are available at `https://github.com/phantatbach/LChange26-Dep`

berg, 2014) or SynGCN and SemGCN (Vashishth et al., 2019). Semantic similarities between words or tokens (in the case of contextualised models) are often calculated using cosine distance between their vectors. For semantic change detection tasks, the standard workflow is as followed (Tahmasebi and Dubossarsky, 2023; Tahmasebi et al., 2021):

1. Data are split into different periods;

2. Semantic vectors of the target words are trained for each time period,;

3. Different embedding spaces are aligned;

4. Semantic changes are quantified by comparing aligned representations, often via cosine distance between word vectors, or via clustering comparison of representations across periods.

While collocation-profile approaches keep the evidence directly interpretable, they often yield descriptive outputs that are harder to compare systematically across datasets and languages. As lexical semantic change (LSC) research moved toward standardised evaluation with different standard shared tasks and dataset, amongst which the most popular one is SemEval 2020 Task 1 (Schlechtweg et al., 2020), methods that produce compact numeric representations and a single comparable distance score became increasingly attractive. As a result, the majority of works and even the current state-of-the-art are all applications of this line of method (e.g., Cassotti et al., 2023; Geeraerts et al., 2023; Giulianelli et al., 2020; Kim et al., 2014; Kulkarni et al., 2015; Kutuzov et al., 2018; Montariol et al., 2021; Tahmasebi and Dubossarsky, 2023; Tahmasebi et al., 2021).

There are a few exceptions to this trend, who did not rely on dense numerical vectors, such as Ryzhova et al. (2021) and Kutuzov et al. (2021) who used the distributions of grammatical features to detect semantic changes and achieved great results, even surpassing many embedding methods in some tasks and datasets; or Tang et al. (2023) who compared sense distributions to detect semantic change. Inspired by their work, we want to investigate the ability of using only explicit linguistic information (in this case, syntactic co-occurrences) in capturing word meaning change.

Most of the work in collocation/co-occurrence analysis uses bag-of-words window co-occurrence

and suffers from arbitrariness in their choice of context windows and from accidental co-occurrences (Evert, 2008). We instead rely on co-occurrence analysis with dependency defined contexts. Rather than using surface window contexts, it treats dependency relations (slots) as interpretable dimensions, so predicted change could be attributed to concrete shifts in those dependency environments. This is conceptually similar to the DepDM model of the Distributional Memory Framework (Baroni and Lenci, 2010) and to the Sketch Engine's Word Sketch Difference (Kilgarriff et al., 2014) but in our case, we keep the raw dependency relations instead of merging them, as will be discussed below. Also, most of them only focus on the salient co-occurrences with such measures as logDice (Rychlý, 2008), MI (Church and Hanks, 1990) or Local MI (Baroni and Lenci, 2010), and LLR (Dunning, 1993), thus do not measure the dynamic of the shift of the co-occurrence profile. We quantify semantic change using Jensen-Shannon Divergence (Menéndez et al., 1997). It is calculated between the distribution of the lexical fillers (slot fillers) of each slot across periods. We evaluate our method in the standard shared tasks, SemEval 2020 Task 1 (Schlechtweg et al., 2020) with 4 languages. Because we use all dependency slots, our method is unsupervised, making it directly comparable to other systems. Finally, we also conduct in-depth qualitative analyses to highlight both the usefulness and limitations of our approach.

3 Data and tasks

Following the SemEval-2020 task on Unsupervised Lexical Semantic Change Detection, we formulate the task as either binary classification or graded ranking, corresponding to Subtask 1 and Subtask 2 in SemEval-2020 (Schlechtweg et al., 2020). In Subtask 1, given a set of target words, a system has to predict whether each word gained or lost at least one sense between two time periods. In Subtask 2, a system has to rank target words by their degrees of semantic change. Annotating data for semantic change is non-trivial since it requires judging meaning differences across many usage instances from each period. A widely adopted annotation protocol is DURel (Schlechtweg et al., 2018), where annotators are asked to compare pairs of sentences containing the target word and judge whether the word is used in the same or a different sense, then these pairwise judgments are aggregated into a final

change score, either binary or continuous. This was the procedure used in the creation of the SemEval 2020 dataset. It covers four languages, namely English (37 targets), German (48), Latin (40), and Swedish (32), with manual annotations for both subtasks. Each language is accompanied by a diachronic corpus split into two time periods. Using that dataset allows us to directly compare our approach against state-of-the-art distributional systems under a shared evaluation setup.

4 Method

4.1 Data preprocessing

Because our method requires dependency relations, which are not available in the original dataset, we first parsed the data using Stanza (Qi et al., 2020) given its high performance on different languages. Upon inspection, we noticed some discrepancies between the Stanza outputs and the lemmatised SemEval data in **part-of-speech (POS) tags** and **lemmas**.

POS mismatches The original dataset did not take into account the POS tags of the target lemmas except for English where targets are provided as lemma_POS combinations (e.g., attack_nn but not attack_vb) because annotation was performed for specific lemma–POS pairs (Schlechtweg et al., 2020). However, because our package requires the POS tags as part of the input, ensuring POS consistency is crucial. We found cases where tokens corresponding to nominal targets were tagged as PROPN by Stanza (causing missed target occurrences if we filter by NOUN), and cases where adjectival targets were tagged as VERB when used as participles.

Lemma mismatch We also found cases where Stanza's lemmatization diverges from the official target lemma due to orthographic variation and OCR errors. One notable example is the German noun *Lyzeum*, which is consistently lemmatised as *Lyceum* by Stanza in the earlier period (old standard spelling), leading to systematic mismatches with the SemEval target form.

To quantify the impact of these discrepancies, we prepared two parallel versions of the parsed data, Stanza-faithful (SZ) and SemEval-faithful (SE). In the SZ version for all languages, we only do minimal normalisation to the target lemmas, namely changing PROPN to NOUN and spelling variations. In the SE version of German, Latin, Swedish, we

ensure that the Stanza results match those of the original lemmatised dataset by replacing all POS tags of the target lemmas with TAR and normalising the spelling variations of the target lemmas. For the English SE version, we split the dataset into multiple sub-datasets based on the official lemmatised files where the targets are marked. Each sub-dataset corresponds to one target lemma.

4.2 Basic procedure

For the main experiment, we used SynFlow (Phan-Tat, 2025), a custom package developed specifically for this method.

For each target lemma (e.g., the English noun *plane*), we extract all of its dependency slots (e.g., adjective modifiers or amod) and the slot-fillers and their POS tags (e.g., new/A, old/A, vertical/A, horizontal/A). Note that because the CoNLL-U file does not encode the directionality in the deprel column but only in the HEAD indices, we have to use the prefix chi_ (children) if the slot-filler depends on the target and the prefix pa_ (parent) if the target depends on the slot-filler. We preserve distinct dependency labels instead of aggregating semantically similar roles (e.g., combining nsubj:pass and obj). Doing so has two advantages. First, this semantic granularity allows us to detect specific structural shifts, such as an increase in passive voice usage, which would otherwise be obscured by a coarse-grained object category. Second, it is generalisable across languages while merging has to be performed on a case-by-case basis and requires language specific knowledge.

For each target word and for each of its slots, we now have a different distribution of their slot-fillers for each period. To quantify how this distribution changes across periods, we use Jensen-Shannon Divergence (JSD). JSD is a symmetric, smoothed variant of Kullback–Leibler divergence (Kullback and Leibler, 1951) and when computed with log base 2, it is bounded between 0 and 1, making the scores easy to interpret and comparable across slots. Moreover, JSD can be decomposed into per-filler contributions, enabling fine-grained attribution of which fillers drive a slot's change. By contrast, prior work on Grammatical Profiles (Kutuzov et al., 2021; Ryzhova et al., 2021) measured change with cosine distance between time-specific frequency vectors, which compares vector geometry (angles) rather than directly comparing probability distribu-

tions.

After computing slot-level JSD scores for each target lemma, we aggregate them into a single lemma-level change score by taking the mean JSD over slots whose JSD exceeds 0.5. We use a mean rather than a sum because lemmas differ in how many slots they have and summing would systematically inflate scores for lemmas with more slots. Using the maximum is also undesirable because 1) very rare slots with small sample size can yield extreme JSD values (often 1), which would make many lemmas appear equally 'maximally changed' and 2) of our assumption that semantic change is manifested in multiple slots. There is little established guidance on choosing an appropriate JSD threshold, and sparsity becomes more severe when distributions are split by slot, which can push JSD values upward compared to larger distributions. The 0.5 value was empirically selected (based on our previous experience applying JSD to sparse linguistic data) to balance sensitivity to genuine linguistic change against robustness to noise. It serves as a high-pass filter for semantic salience, which isolates substantial structural shifts while suppressing minor fluctuations inherent to sparse distributions. Note that we intentionally avoid any supervised finetuning (e.g., for selecting the optimal JSD cut-off) to preserve the fully unsupervised protocol and comparability with other unsupervised approaches.

To sum up, the inputs to our method are dependency-parsed corpora and a list of target lemmas. SynFlow then

1. Collects the slot fillers for each dependency slot;

2. Computes the JSD scores for every slot;

3. Aggregates the slot-level JSDs greater than 0.5 into a lemma-level JSD score.

The output is a dictionary of the form {`target lemma: aggregate JSD score`}. In the next sections, we describe 2 improvements to this basic procedure.

4.3 Frequency filtering

To reduce noise that could be introduced by rare items, we exclude slot-fillers that appear only once in all periods before calculating the JSD.

4.4 POS removal

In the basic procedure above, we initially retain POS tags for slot fillers. However, because our analysis is driven primarily by dependency relations, and because POS tags can introduce additional noise (e.g., adjective vs. participial verb ambiguities as discussed above), we remove the POS information from slot-fillers before computing JSD. This reduces the risk that parser-induced POS inconsistencies create spurious differences between periods and ensures that divergence reflects changes in dependency defined co-occurrence rather than tagging artefacts.

5 Results

We evaluate our method on both subtasks of the SemEval 2020 Unsupervised Lexical Semantic Change Detection shared task. As described in Section 3, Subtask 1 is a binary classification task and is evaluated based on accuracy. Subtask 2 is a ranking task and is evaluated with Spearman's rank correlation. The main focus of our method would be on **subtask 2**. We then use our subtask 2 scores for subtask 1 classification, following the strategy of Kutuzov et al. (2021).

5.1 Subtask 2

The results for subtask 2 are reported in table 1. For a full comparison with SemEval-2020 Task 1 participating systems, see Table 5 in Schlechtweg et al. (2020).

Basic procedure The slot-level JSD aggregation described above yields an average Spearman correlation of 0.096 on the SE dataset and 0.108 on the SZ dataset. This is below the shared-task baseline, but still higher than some systems based on type and token embeddings (e.g., UoB, TUE).

Frequency filtering Removing rare slot-fillers (Section 4.3) improves the performance substantially, especially on the SE dataset: the average correlation across languages nearly doubles, and Latin improves by an order of magnitude ($0.026 \rightarrow 0.215$). On the SZ dataset, the effect is mixed: English drops to -0.001, while the other three languages improve.

POS Removal Removing POS tags from slot-fillers yields only a small additional gain: $+0.021$ on SE and $+0.005$ on SZ on average.

System	Average	English	German	Latin	Swedish
Basic Procedure					
SE	0.096	0.108	0.135	0.026	0.115
SZ	0.108	0.139	0.120	0.017	0.155
Frequency Filtering					
SE 2	0.229	0.266	0.253	0.215	0.181
SZ 2	0.129	-0.010	0.191	0.142	**0.192**
POS Removal					
SE No POS	0.117	0.115	0.175	0.038	0.141
SZ No POS	0.113	0.139	0.145	0.001	0.166
Frequency Filtering + POS Removal					
SE 2 No POS	**0.239**	**0.277**	**0.258**	**0.258**	0.162
SZ 2 No POS	0.139	-0.008	0.193	0.197	0.173
Prior SemEval results					
Count baseline	0.144	0.022	0.216	0.359	-0.022
Best shared task system (Cassotti et al., 2023)	0.583	0.757	0.877	-0.056	0.754

Table 1: Performance in graded change detection (SemEval 2020 Subtask 2), Spearman rank correlation coefficients. SE = SemEval, SZ = Stanza.

System	Average	English	German	Latin	Swedish
Basic Procedure					
SE	0.580	0.595	0.604	0.475	0.645
SZ	0.583	0.541	0.604	0.475	**0.71**
Frequency Filtering					
SE 2	0.593	**0.649**	0.604	0.475	0.645
SZ 2	0.569	0.486	0.604	0.475	**0.71**
POS Removal					
SE No POS	0.59	0.595	**0.646**	0.475	0.645
SZ No POS	0.583	0.541	0.604	0.475	**0.71**
Frequency Filtering + POS Removal					
SE 2 No POS	**0.606**	**0.649**	0.604	**0.525**	0.645
SE 2 No POS DP	0.585	0.621	0.604	0.5	0.613
SZ 2 No POS	0.565	0.486	0.604	**0.525**	0.645
Prior SemEval results					
Count baseline	0.613	0.595	0.688	0.525	0.645
Best shared task system	0.687	0.622	0.75	0.7	0.677

Table 2: Performance in binary change detection (SemEval 2020 Subtask 1), accuracy. Note that in this paper we mostly focus on ranking (Subtask 2). All the binary change detection methods here are entirely based on the scores produced by the ranking methods. SE = SemEval, SZ = Stanza.

Frequency Filtering + POS Removal   Combining both strategies produces our strongest overall results. Performance improves for all languages except Swedish (where the combined setting reaches 0.162, slightly below our Swedish best of 0.192). The combined setting on the SE dataset achieves an average Spearman correlation of 0.239, which is nearly double the Count baseline and higher than

a number of embedding-based systems. This is impressive given the simplicity of the approach.

5.2   Subtask 1

The results for subtask 1 are presented in table 2. We follow the thresholding strategy used by Kutuzov et al. (2021) by assigning a classification score of 1 to the top 43% of the target words for

each language. The ranking is obtained from their JSD scores as discussed above. The different average accuracy scores from the different systems are tightly clustered (between 0.58 and 0.606) and are very close to the accuracy of the baseline, which is already surprisingly hard to beat.

Our best system is still SE 2 No POS (i.e., SemEval + frequency filtering + POS removal), which achieves an average accuracy of 0.606, showing that simple denoising techniques (i.e., frequency filtering and POS removal) work. In the case of Swedish, 3 of our system variants even outperform the state-of-the-art, achieving an accuracy of 0.71. We also tested our best system with dynamic programming[2] (Truong et al., 2020) and observed that it does not cause any significant accuracy decrease when compared to the hard-coded 43% cut-off point. This indicates that our method does not require knowledge of the test data distribution

6 Qualitative analysis

The strong point of our system is its interpretability, in the sense that it allows the users to see which kinds of shift drive the change through JSD decomposition, and check the concordances for the locus of semantic change. In this section, we analyse the prediction of the system qualitatively to understand what is being captured.

Based on the rankings of the Subtask 2 Gold Scores and our system predictions, we divided the target lemmas into five groups: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) and MID. A lemma is TP if it falls in the top 33% of both the gold and predicted rankings, and TN if it falls in the bottom 33% of both. A lemma is FP if it is in the bottom 33% of the gold ranking but the top 33% of our method's ranking, and FN if it is in the top 33% of the gold ranking but the bottom 33% of our method's ranking. All remaining lemmas are assigned to the MID group. We focus our qualitative analysis on TP, TN, FP, and FN, as these categories most clearly reveal our method's strengths and limitations. Due to space constraints, in the main text we present one illustrative example per group and language (the most extreme cases). The lemmas of each group are listed in Table 3 in the Appendix.

Also, in the analysis, we will only look at slots with JSD of at least 0.5 as that is the cut-off point used in the previous procedure. A detailed description of the dependency slots presented in this section is provided in Table 4 in the Appendix.

True positives Upon qualitative analysis, our method does indeed capture relevant co-occurrence information related to semantic change. The English noun *graft* develops a new MEDICAL use (transplanting tissue) and a new POLITICAL use (political corruption) in the later period, whereas in the earlier period it is used primarily in agricultural contexts. This is reflected in 3 out of 5 dependency slots, such as governing verbs (pa_obj, e.g., *cut the graft*) and compound modifiers (chi_compound, e.g., *skin graft*) as shown in Figure 2. In the first period, fillers and contexts predominantly relate to cultivation and horticulture (e.g., *adapting the graft to the soil*, *flowering graft*), while the second period is characterised by an increase of medical and political contexts (e.g., *help the skin graft*, *government graft*). The remaining two slots show weaker or less interpretable shifts. The German adjective *abgebrüht* has developed a new FIGURATIVE meaning (e.g., *abgebrühter Manager* 'a hardened manager') from the LITERAL one (e.g., *abgebrühter Reis* 'blanched rice'), as evidenced in the modified nouns (pa_amod) and modifying adverbs (chi_advmod). The Latin noun *pontifex* shifts from 'state-religious officer' toward 'the Pope'. This is reflected in 13/15 slots. In the adjectival modifier (chi_amod) and oblique nominals (pa_obl) slots, for example, *pontifex* co-occurs less with administrative vocabulary (e.g., *gero* 'manage') and
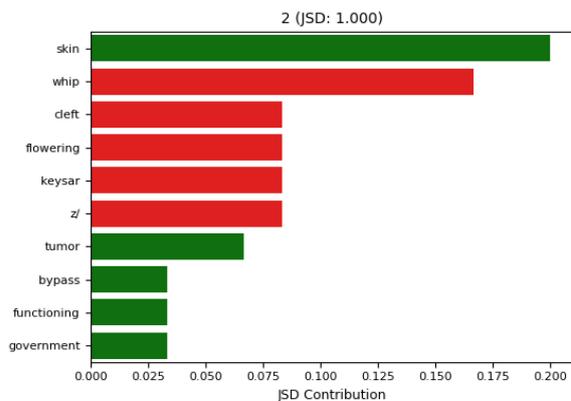

Figure 2: The JSD contributions of different slot-fillers in the slot chi_compound of the English noun *graft*. Green bars indicate an increase in relative frequency; red bars indicate a decrease.

---

more with Christian contexts (e.g., *summus pontifex* 'Supreme Pontiff', *oro* 'pray'). Finally, the Swedish noun *uppläggning* has developed a new abstract STRUCTURING meaning from the physical STOCKPILING meaning, with consistent evidence in all of its slots. For example, in the nominal modifier slot (`pa_obl`), there is an increase in *räkenskap* 'accounting' while in the earlier period it is dominated by types of goods such as *jernplätars* 'of iron plates', *spannmåls* 'of cereals'.

True negatives Regarding TNs, we observe that few dependency slots exceed the JSD threshold and even when they do, the margin is often negligible. This is the case for the English noun *tree*. Our method detects a moderate shift in nominal modifiers (`chi_nmod`, JSD = 0.517) and compounds (JSD = 0.533), but these shifts are driven by a transition from biological attributes (e.g., *trees of this size*) to locative contexts (e.g., *trees of this yard*). Crucially, however, both filler sets remain semantically consistent with the core PLANT sense, indicating mere topical variation rather than genuine semantic evolution. Even if there are slots with high divergence (e.g., ≈ 0.8), the aggregate lemma-level score remains low because other stable slots dilute the overall mean. This was the case for the German verb *vergönnen*, the Latin adjective *necessarius* and the Swedish verb *uträtta*.

False positives FPs are target words for which our method detects large shifts in one or more slots, yet the lemma is ranked low in the gold Subtask 2 ranking. A common pattern is slot-level turnover driven by a small number of highly characteristic contexts, which inflates divergence without corresponding to a clear sense innovation under the annotation scheme. For the English noun *face*, the high change score is driven primarily by the two compound slots (`pa_compound`, `chi_compound`). The divergence is dominated by a few fixed expressions, most notably *face value* and product-related compounds such as *face cream* and *face lift*. This illustrates the limitation where the substantial frequency change of a small set of expressions can cause the method to overestimate the semantic change. The Swedish noun *kokärt* is also stable in meaning, yet all of its slots exhibit high JSD due to contextual turnover. The German noun *Seminar* presents an interesting case. We observe a metonymic shift: INSTITUTIONAL DEPARTMENT → COURSES. This is strongly evidenced by a high divergence (0.926) in

the nominal heads (`pa_nmod`), transitioning from institution-related fillers (e.g., *Lehrer*, *Director*, *Übungsschule*) to participant-oriented ones (e.g., *Teilnehmer*). However, its official gold label for Subtask 1 is 0 (i.e., no sense gain or loss) and its official Subtask 2 rank is only 39/48 (i.e., low degree of semantic change). This suggests our system successfully detected an overlooked semantic shift in the official annotations. The Latin noun *dolus* is another interesting case. It is officially classified as 1 in Subtask 1 and our method successfully identified the word's narrowing and pejoration from TRICK (neutral) to FRAUD (negative) in the governing verb (`pa_obj`) slot (JSD = 0.933). However, its official ranking in Subtask 2 is low (29/40). This strengthens the organisers' finding: models that excel at one subtask are not necessarily good at the other (Schlechtweg et al., 2020).

False negatives FNs are target words whose aggregate JSD scores underestimate the degree of change with respect to the gold ranking. We attribute a substantial portion of these errors to our slot-selection threshold. Although the English noun *record* develops a salient BEST PERFORMANCE sense (e.g., *world record*), reflected by a sharp rise of *set* and *break* in governing verbs (`pa_obj`), the divergence for this slot is 0.48, just below our 0.5 cutoff. This signal is thus excluded from aggregation, causing the lemma-level score to miss the underlying shift. A similar threshold effect appears for the Swedish noun *motiv*, which shifts from a predominantly REASON meaning (typically framed in moral evaluation) to a predominantly ARTISTIC SUBJECT meaning (often framed in aesthetic evaluation) between the two periods. While detectable in adjectival modifiers (`chi_amod`), the slot's low JSD (0.436) leads to its exclusion, again resulting in an underestimated aggregate score. In other cases, the change signal is present but remains only moderately strong at the slot level, so it does not lift the aggregate score enough to match the gold ranking. The German verb *ausspannen* shows a clear metaphorisation: PHYSICAL (unharnessing/stretching out) → RELAX (taking time off). This is visible in its direct objects (`chi_obj`) with an increase in temporal fillers (*Tag*, *Stunde*) and a decrease in physical objects (*Draht*, *Faden*), or in the passive subject (`chi_nsubj:pass`). However, their JSDs are only 0.543 and 0.584, respectively. This illustrates that genuine semantic innovations can yield relatively modest divergences in individ-

ual slots, resulting in underestimated lemma-level change. A related pattern occurs for the Latin noun *potestas*, where broadening (LEGAL AUTHORITY → GENERAL CAPABILITY) is reflected in two low JSD slots, adjectival modifiers (`chi_amod`, JSD = 0.587) and nominal modifiers (`chi_nmod`, JSD = 0.643).

7 Conclusion

This paper demonstrates that explicit linguistic co-occurrence evidence, when structured through syntax, is sufficient to build an effective lexical semantic change detection system. Our method achieves competitive performance across languages and can even outperform several distributional models in some tasks and languages. Although our method does not match state-of-the-art systems based on large pretrained language models, which can encode a wider range of linguistic and contextual information, maximising performance metrics was not our primary objective. Instead, we want to bring more attention to theory-driven and interpretable approaches.

To further refine this approach, future work should:

1. Replace hard slot thresholds with more adaptive, evidence-weighted aggregation to increase reliability;

2. Further develop theory-driven ways of selecting and merging slots to reduce non-essential variations (recall the `nsubj:pass` and `obj` example in 4.2), but note the cost of generalisability;

3. Explore hybrid methods that preserve slot-level attribution while using embeddings to stabilise sparse evidence (e.g., clustering slot-fillers using general type embeddings before computing JSD);

4. Take into account covariations between slots and looking at multiple slots simultaneously (e.g., to detect changes in valency/subcategorisation patterns)

More broadly, we view our method as an instance of a general framework for semantic change analysis: identify interpretable dimensions of a target's distribution (here, dependency relations) and quantify change within each dimension before aggregating into a global measure. This framework could (and should) be used as a complementary tool to interpret and validate the changes detected by other approaches.

8 Limitations

We highlight three primary limitations and practical considerations of our proposed method.

First, dependency availability is a hard constraint, making this method inapplicable to low-resource languages lacking robust dependency parsers. Even for high-resource languages, preprocessing choices matter: POS and lemma mismatches between gold targets and parsers (e.g., Stanza) output requires alignments.

Second, sensitivity to noise (e.g., rare slots, rare slot-fillers) remains a challenge. Denoising (e.g., filtering singletons and removing POS tags from fillers) proved central to our performance.

Third, our qualitative analysis, while supporting the interpretability claim, simultaneously reveals inherent constraints in our scoring mechanism. The method can *overestimate* change when a small number of fixed expressions or highly productive co-occurrences dominate slot turnover (as seen with idioms). Conversely, it can *underestimate* change when genuine semantic innovations fall just below the heuristic slot-selection cutoff.

Acknowledgments

References

Josep Alba-Salas. 2007. On the life and death of a collocation: A corpus-based diachronic study of dar miedo/hacer miedo-type structures in Spanish. *Diachronica*, 24(2):207–252. Publisher: John Benjamins Type: Journal Article.

Laurence Anthony. 2014. AntConc (Version 3.4. 3)[Computer Software]. Tokyo, Japan: Waseda University.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and

Francisco Herrera. 2019. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146. Place: Cambridge, MA Publisher: MIT Press.

Vaclav Brezina, Pauline Weill-Tessier, and Tony McEnery. 2021. #lancsbox v. 6.x. Lancaster University. Software version 6.x.

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC Pretrained Model for Cross-Lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Thomai Dalpanagioti. 2019. From corpus usages to cognitively informed dictionary senses. *Lexicography*, 6(2):75–104.

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.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74. Place: Cambridge, MA Publisher: MIT Press.

Stephanie Evert. 2008. Corpora and collocations. In Kytö M Lüdeling A, editor, *Corpus Linguistics. An International Handbook*, pages 1212–1248. Mouton de Gruyter, Berlin, New York.

John Rupert Firth. 1957. A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*, pages 1–32. Basil Blackwell, Oxford.

Marcos Garcia and Marcos García Salido. 2019. A Method to Automatically Identify Diachronic Variation in Collocations. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 71–80, Florence, Italy. Association for Computational Linguistics.

Dirk Geeraerts. 1997. *Diachronic Prototype Semantics: A Contribution Historical Lexicology*. Oxford University Press.

Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang. 2023. *Lexical Variation and Change: A Distributional Semantic Approach*, 1 edition. Oxford University Press.

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, Edinburgh, UK. Association for Computational Linguistics.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. The MIT Press.

Zellig S. Harris. 1954. Distributional Structure. *WORD*, 10(2-3):146–162. Publisher: Routledge.

Bryan Jurish. 2015. DiaCollo: On the trail of diachronic collocations. In *CLARIN Annual Conference 2015 (Wroław, Poland, October 15–17 2015)*, pages 28–31.

Bryan Jurish and Maret Nieländer. 2020. Using DiaCollo for Historical Research. In *Selected Papers from the CLARIN Annual Conference 2019*, pages 33–40.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1:7–36. Publisher: Springer.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages

61–65, Baltimore, MD, USA. 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, CHE. International World Wide Web Conferences Steering Committee. Event-place: Florence, Italy.

S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Andrey Kutuzov, Lidia Pivovarova, and Mario Giulianelli. 2021. Grammatical Profiling for Semantic Change Detection. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 423–434, Online. Association for Computational Linguistics.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Neural Information Processing Systems*.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8):1–42.

Tony McEnery, Vaclav Brezina, and Helen Baker. 2019. Usage Fluctuation Analysis: A new way of analysing shifts in historical discourse. *International Journal of Corpus Linguistics*, pages 413–444.

M.L. Menéndez, J.A. Pardo, L. Pardo, and M.C. Pardo. 1997. The Jensen-Shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. ArXiv:1301.3781 [cs].

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.

Alan Partington. 1998. *Patterns and Meanings*. John Benjamins.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. 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, New Orleans, Louisiana. Association for Computational Linguistics.

Daniela Pettersson-Traba. 2021. Measuring semantic distance across time. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 6(2):138–165.

Bach Phan-Tat. 2025. Synflow.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. ArXiv:2003.07082 [cs].

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Pavel Rychlý. 2008. A Lexicographer-Friendly Association Score. *Proc. 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN*, 2:6–9. Publisher: Masaryk University.

Anastasiia Ryzhova, Daria Ryzhova, and Ilya Sochenkov. 2021. Detection of semantic changes in russian nouns with distributional models and grammatical features. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2021"*.

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, 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)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Hinrich Schütze. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–123. Place: Cambridge, MA Publisher: MIT Press.

V. Seretan. 2011. *Syntax-Based Collocation Extraction*. Text, Speech and Language Technology. Springer Netherlands.

John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

Michael Stubbs. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, 2(1):23–55. Publisher: John Benjamins Type: Journal Article.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. In *Computational approaches to semantic change*. Zenodo.

Nina Tahmasebi and Haim Dubossarsky. 2023. Computational modeling of semantic change.

Xiaohang Tang, Yi Zhou, Taichi Aida, Procheta Sen, and Danushka Bollegala. 2023. Can word sense distribution detect semantic changes of words? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3575–3590, Singapore. Association for Computational Linguistics.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. 2020. Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

Shikhar Vashishth, Manik Bhandari, Prateek Yadav, Piyush Rai, Chiranjib Bhattacharyya, and Partha Talukdar. 2019. Incorporating syntactic and semantic information in word embeddings using graph convolutional networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3308–3318, Florence, Italy. Association for Computational Linguistics.

Appendix

A   Predictions of our method

	English	German	Latin	Swedish
True Positive	graft (noun)	abgebrüht	pontifex	uppläggning
	plane (noun)	Engpaß	sacramentum	krita
	head (noun)	Ohrwurm	scriptura	medium
	bit (noun)	Abgesang	titulus	konduktör
	player (noun)	Eintagsfliege	imperator	uppfattning
		Dynamik	sanctus	beredning
		Armenhaus		notis
True Negative	tree (noun)	vergönnen	necessarius	uträtta
	risk (noun)	Tier	consilium	annandag
	multitude (noun)	Ackergerät	hostis	studie
	fiction (noun)	Frechheit	simplex	
	contemplation (noun)	aufrechterhalten	voluntas	
	bag (noun)		poena	
False Positive	face (noun)	Seminar	dolus	kokärt
	quilt (noun)	Truppenteil	fidelis	vegetation
	lane (noun)	Mulatte	consul	
	savage (noun)	Pachtzins	ancilla	
		Naturschönheit	acerbus	
False Negative	record (noun)	ausspannen	potestas	motiv
	ounce (noun)	artikulieren	licet	ledning
	rag (noun)	verbauen	salus	bearbeta
			virtus	

Table 3: True Positive, True Negative, False Positive, False Negative results of SE 2 NoPOS.

B   Descriptions of the slots

Slot Name	Explanation
chi_advmod	The target is modified by the adverbial slot filler.
chi_amod	The target is modified by the adjectival slot-filler.
chi_case	The target is governing the prepositional slot-filler.
chi_compound	The target is being modified by the slot-filler in the compound.
chi_nmod	The target is modified by the nominal slot-filler.
chi_nsubj:pass	The target is the verb governing the passive subject.
chi_obj	The target is the verb governing the object slot-filler.
pa_amod	The target is the adjective modifying the slot-filler.
pa_compound	The target modifies the slot-filler in the compound.
pa_nmod	The target is the nominal modifier of the slot-filler.
pa_obj	The target is the object of the verbal slot-filler.
pa_obl	The target is the oblique of the slot-filler.

Table 4: A detailed description of the dependency slots presented in section 6.