ConShift: Sense-based Language Variation Analysis using Flexible Alignment

Clare Arrington[1], Maurício Gruppi[2], Sibel Adalı[1]
[1]Rensselaer Polytechnic Institute, Troy, NY, USA
[2]Villanova University, Villanova, PA, USA
arrinj@rpi.edu, mgouveag@villanova.edu, adalis@rpi.edu

Abstract

We introduce ConShift, a family of alignment-based algorithms that enable semantic variation analysis at the sense-level. Using independent senses of words induced from the context of tokens in two corpora, sense-enriched word embeddings are aligned using self-supervision and a flexible matching mechanism. This approach makes it possible to test for multiple sense-level language variations such as sense gain/presence, loss/absence and broadening/narrowing, while providing explanation of the changes through visualization of related concepts. We illustrate the utility of the method with sense- and word-level semantic shift detection results for multiple evaluation datasets in diachronic settings and dialect variation in the synchronic setting.

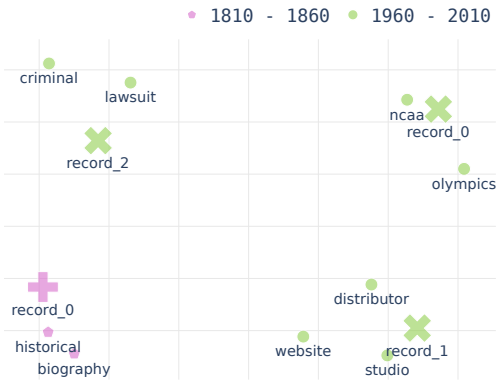

Figure 1: PCA projection of the aligned word embedding space of the 1800s and 2000s from CCOHA. Senses of the word *record* were induced independently and matched following alignment. The single sense from the 1800s (plus sign record_0) broadens into 3 senses with differing contexts in the 2000s.

1 Introduction

We present a series of methods to analyze semantic variation of words across two corpora both qualitatively and quantitatively based on the underlying senses of the words (Kutuzov et al., 2018; Tahmasebi et al., 2021). Word-level semantic change detection methods based on embedding alignment are able to capture the shift in a word's meaning by the distance between the embedded corpora (Hamilton et al., 2016; Gruppi et al., 2021). However, these methods are unable to provide nuanced explanations for the change. More recently, contextualised embeddings have been used for semantic change detection, allowing for sense-level analysis (Montanelli and Periti, 2023). This analysis is still shallow, with few existing methods exploring the relationship between senses from an explainable perspective, such as identifying how new senses arise by extracting words that relate to it (Mitra et al., 2014; Hu et al., 2019; Giulianelli et al., 2020). In part, this is due to the limited evaluation datasets for both word-level (Gulordava and Baroni, 2011; Schlechtweg et al., 2021) and sense-level semantic change detection (Zamora-Reina et al., 2022). Thus, methods often focus on predicting whether a word has distinct meanings across corpora or not, as opposed to finding partial overlap or the existence of multiple senses of the same lexeme. Certain sense-level changes in one corpus may not be easy to capture in these methods as analysis often looks at senses in both corpora jointly.

We introduce a set of methods for sense-aware detection of semantic variation, called ConShift, leveraging the local information provided by contextual representation of LMs and the interpretability of distributional semantics methods leveraged through alignment. Through this, we detect changes in senses between two corpora and visualize them through their local contextual neighbors. Figure 1 provides an example for the word *record* from the Clean Corpus of Historical American English (CCOHA) (Alatrash et al., 2020). With a single sense in the 1800s, *record* acquires multiple new, related senses in the 20th century, indicating a process of semantic broadening. Our visualiza-

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 167–181

tion shows not only these individual meanings but exposes their semantic relationship across corpora on a common axis with the help of an alignment algorithm. ConShift can be used to extract senses of a word in diachronic (over time) and synchronic (same time, different domains) settings, and illustrate the semantic relationship between senses even when such differences are complex.

Our work makes the following contributions: (1) We present ConShift, a family of methods for aligning sense-aware word embeddings across a pair of input corpora to analyze different usages of a word. This method captures word-level semantic change, as well as sense loss, gain, narrowing and broadening. (2) We show our methods work in both diachronic and synchronic settings through multiple datasets, even for senses of words not used in the alignment step. (3) Our methods handle multiple forms of word sense assignment and clustering. We show an analysis of the impact of both, especially clustering of senses jointly and independently across two corpora. (4) We develop explanation methods that can be used to visualize senses in a common axis through PCA projection after alignment, capturing both related terms and semantic distances. The code and embeddings are available at https://github.com/clare-arrington/ConShift.

2 Related Work

Sense-based Semantic Change Modeling The primary task in sense change modeling is unsupervised diachronic shift detection, where the goal is to identify changes in a word across two temporal corpora (Kutuzov et al., 2018; Tahmasebi et al., 2021). Recent methods for this task have focused on contextualised architectures like BERT (Devlin et al., 2018) and XLM (Conneau and Lample, 2019) for dynamic representations of a word's usage in context. Montanelli and Periti (2023) organized contextualised semantic shift detection methods into a framework based on meaning representation, time-awareness, and learning modality. At a high level, contextual representations are assessed by directly comparing embeddings (Cassotti et al., 2023; Zhou and Li, 2020; Kutuzov, 2020; Rosin and Radinsky, 2022; Pömsl and Lyapin, 2020), or by clustering into senses before detection (Hu et al., 2019; Giulianelli et al., 2020; Montariol et al., 2021; Arefyev and Zhikov, 2020; Periti et al., 2022).

When sense-based methods use clustering for consecutive time periods, the resulting clusters need to be matched on similar word meanings before detection. Kanjirangat et al. (2020) considered 2 approaches: joint clustering and matching cluster centers by Euclidean distance to create one-to-one pairs. Tang et al. (2023) annotated senses by comparing contextual embeddings of usages in a corpus to sense embeddings. Alternatively, some methods want to capture diachronic shift over a smooth time period (Frermann and Lapata, 2016). Periti et al. (2022) offered an alternate approach via incremental clustering, where the corpora at different time steps is progressively split into senses clusters. Montariol et al. (2021) skipped cluster alignment by using the Wasserstein distance (Solomon, 2018), an optimal transport problem that finds the minimal effort needed to transform one distribution into another. Both methods show distribution-based changes to the senses over time; they did not capture the relationship between senses. In contrast, our method uses embedding alignment, which allows for comparisons between all senses of a word.

One recent approach for sense modeling uses word substitutions to observe semantic change. Kudisov and Arefyev (2022) (BOS) induced explainable senses through Masked Language Modeling (MLM), to generate lexical substitute vectors (Amrami and Goldberg, 2019). These vectors provided more information than contextualised embedding layers, since sense clusters can be described by their highest lexical substitutes. Card (2023) simplified the form of vector representation from MLM and used Jensen-Shannon divergence to measure between target probability distributions, while accounting for frequency. Periti et al. (2024) used a replacement schema to simulate forms of change, extending to additional models including LLaMa 2. The use of a larger model showed improvements over the smaller LMs like BERT. These approaches provide interpretable process, though there is room for further explainability once semantic shift has been detected. Giulianelli et al. (2020) tracked senses over time, through k-means clustering on BERT embeddings. They provided manual observations of cases like narrowing and broadening, and distinctions between sense clusters like metaphorical usage. Hu et al. (2019) used a supervised sense disambiguation approach to match senses from the Oxford English Dictionary to word usages and tracked frequency over time. They performed an in-depth, manual modeling of temporal change in individual senses through an ecological perspective. ConShift is able to detect and label

cases of sense change, like narrowing.

Semantic Change Detection through Alignment Word embedding alignment is used to measure word distances across multiple embeddings for semantic shift detection (Kim et al., 2014; Kulkarni et al., 2015; Bamler and Mandt, 2017; Basile et al., 2020). Hamilton et al. (2016) is a projection-based approach that uses Orthogonal Procrustes to set one word embedding vector in the same space as another through the use of landmark terms which are chosen as the subset used in alignment. Some approaches for aligning embeddings select terms based on their presence in both corpora (Yin et al., 2018), their frequency, or their stability across the two vectors (Lubin et al., 2019; Gruppi et al., 2021).

3 Methodology

3.1 Overview

In this section, we introduce our approach to sense-based semantic variation analysis. Given two corpora of the same language, C_1 and C_2, and a set T of target words, we produce quantitative scores and qualitative results for analyzing the semantic variation. Corpora C_1 and C_2 may originate from different domains over same time period (synchronic variation) or from the same domain over different periods of time (diachronic variation). The outline of the method is as follows:

1. For each target word $t \in T$, we extract senses. The base method we use is based on clustering of masked token predictions of a transformer model (using MLM), grouping tokens with similar contexts to the same cluster/sense. Instances of target tokens in each corpus are tagged with the sense cluster they belong to.

2. Sense-enriched static word embeddings are trained for each corpus, where sense-tagged tokens are embedded as new types. The embeddings are aligned with a self-supervised matching method to compare semantic spaces.

3. After alignment, sense-tagged types are matched to detect whether a given sense of a word in C_1 is present in C_2 and vice-versa. If senses can be matched, i.e. are within a given cosine distance after alignment, that meaning exists in both corpora. Otherwise, a sense only in one corpus implies a semantic variation in the word's usage between two corpora.

We introduce three variations of the ConShift method. In ConShift and ConShift-M-J, all usages of a target word from both corpora are clustered jointly into senses. In these approaches, clustering provides an initial pairing between senses across the two corpora once embedded. In contrast, target usages are clustered independently for each corpus in ConShift-M-I ,resulting in senses that are unmatched initially. We test the semantic shift of paired senses and their target words through alignment of corpora. While ConShift only checks for shift between one to one sense matches, ConShift-M-J and ConShift-M-I both allow checking of many-to-many (0..M, M..M or M..0) matches. Figure 2 shows the overview of ConShift and the following sections describe the details of each step.

3.2 Obtaining Sense-Tagged Tokens

The starting point of our approach is the assignment of senses to occurrences of target words in each corpus through word sense disambiguation (WSD). Given a set of sentences, or usages, that contain a target word, we cluster them into senses. Individual target occurrences are then tagged with a suffix label indicating their cluster and representing a unique sense token. For example, if the word *staff* was found to have two senses, then token *staff* becomes *staff_1* or *staff_2* for all instances within the corpora. When joint clustering is used for inducing senses, labels in both corpora will refer to the same set of clusters, meaning *staff_1* from C_1 is related to *staff_1* in C_2. With independent induction, these labels are unrelated and must be matched during or after alignment.

For sense assignment, any well-established WSD algorithm can be used, including supervised methods that assign senses to an existing inventory or unsupervised methods, i.e. word sense induction (WSI). We select a method that uses Masked Language Modeling (MLM) with hierarchical agglomerative clustering (HAC) as the base for all results presented in the paper, as one of the best performing methods in the literature for WSI (Amrami and Goldberg, 2019). In Appendix Section A.1, we compare results against an alternate method using k-means (Giulianelli et al., 2020) and present full implementation details of these methods for completeness and reproducibility.

Through BERT's MLM task (Devlin et al., 2018), the token of a target word in a sentence is masked and the MLM predicts which terms should occupy the masked position. This prediction results in a

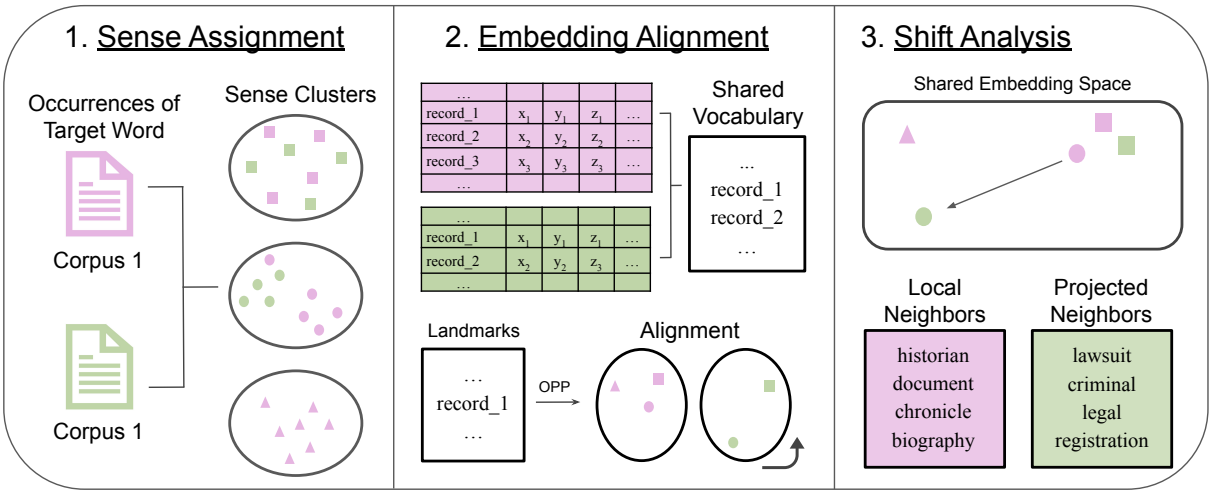

Figure 2: Overview of ConShift for a pair of input corpora and target word, *record*. Embedded occurrences are clustered jointly into senses. After tagging, corpora are transformed into separate word embeddings and aligned. Semantic variation is detected by measuring distance between shared senses.

probability distribution over the entire vocabulary of the language model, with higher probability indicating a good lexical replacement for the missing token. This process transforms sentences from the corpus into lexical substitute vectors, which are then clustered with HAC, representing a subset of word usages from the corpus. We use these clusters to transform the corpora by replacing the target tokens with sense-labeled tokens.

3.3 Aligning Sense-Tagged Word Vectors

After replacing every target term, we use the sense-embedded corpus to create static word embeddings with Word2Vec (Mikolov et al., 2013). This procedure generates two embedding matrices E_1 and E_2, corresponding to corpora C_1 and C_2. We align E_1 and E_2 to measure the distances of sense-tagged word vectors across embeddings. Alignment is necessary when comparing independently-trained word embeddings due to the stochastic nature of the initialization of the model's weights, which renders a direct comparison impossible. To perform alignment, we find an optimal transformation that maps the semantic space of one embedding to the other that minimizes the distance between a set of anchor words, or *landmarks*. Since this process hinges on landmark tokens being relatively unshifted, selecting the landmark subset from the vocabulary is a non-trivial and crucial step for effective semantic shift detection (Yin et al., 2018; Lubin et al., 2019; Gruppi et al., 2021).

3.3.1 Choosing Sense-Aware Landmarks

Landmarks are by definition stable (unshifted) tokens. As these unshifted tokens are not known apriori, we select the set using the search method from S4-A[1] (Gruppi et al., 2021). The method employs a self-supervised approach to identify term vectors that are shifted between embeddings E_1 and E_2. On the initial iteration, tokens from the entire shared vocabulary are used for global alignment. The vocabulary is bisected into landmarks and non-landmarks based on ranked cosine distance. A neural network classifier is then iteratively trained using the stable landmark set and a set of vectors that have been artificially shifted to be unstable. After each iteration, the classifier updates its weights and relabels the original term vectors into landmarks and non-landmarks. Using the new landmark set, the embeddings are aligned once more. This procedure repeats until no change occurs from one iteration to the next, defining a consistent landmark set.

Because our embeddings contain both word and sense tokens, we need an additional step to prepare the shared vocabulary, i.e. the set of tokens that exist in both E_1 and E_2. When all tokens represent words, the shared vocabulary is the intersection of V_1 and V_2 and words are paired trivially. Similarly, when we assign the same set of senses to both embeddings, either through WSD or joint WSI, sense tokens are also paired together through the intersection of the vocabularies. Effectively, we treat these senses as regular word tokens during alignment,

[1] https://github.com/IBM/S4_semantic_shift

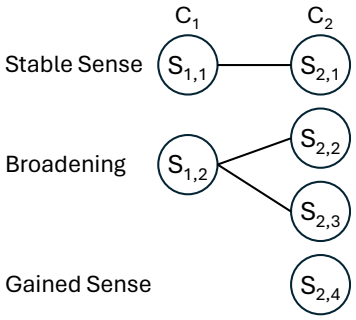

Figure 3: Example matching of a set of senses for diachronic shift detection, where corpus C_1 is older than C_2. A sense in C_1 with 2 or more matched senses in C_2 represents broadening of usage, while the inverse represents narrowing. Lost and gained senses are those without any matches.

since they represent a common cluster origin. This approach is used for the base method, ConShift, where we jointly cluster word usages from the two input corpora. Hence, the same token in E_1 and E_2 represents all sentences corresponding to a single sense cluster (i.e. *staff_1* in E_1 is equivalent to *staff_1* in E_2.) In this case, quality of landmarks is limited by the quality of the initial clusters.

To account for the interrelation of senses, our second approach considers that additional matches may exist between senses. We expand the shared vocabulary by adding pairs consisting of all senses of a target in both E_1 and E_2. Given senses $s_{1,1}, ..., s_{1,n}$ in corpus E_1, and senses $s_{2,1}, ..., s_{2,m}$ in corpus E_2, we generate the following vector tuples: $(s_{1,1}, s_{2,1}), (s_{1,1}, s_{2,2}), ..., (s_{1,n}, s_{2,m})$. Thus, the representation of the original target word is fully covered in the alignment process. When searching the shared vocabulary, zero, one or many pairs containing a singular sense may be selected as landmarks. This approach may be applied to any form of sense assignment, shared or not. In ConShift-M-J, we cluster jointly and pair all senses of a target regardless of whether their token is shared, as opposed to ConShift. In ConShift-M-I, clusters are induced independently to be optimized for each corpus. No information is known about the relationship between senses prior to pairing. However, these pairings can now be used as landmarks during the alignment process. Being selected as a landmark can indicate at a high-level if a token is stable or not, but we don't rely on this information for shift detection or to filter sense pairings. After alignment, we measure and evaluate all sense pairings fully.

3.3.2 Alignment

We align the embedding matrices through the Orthogonal Procrustes (OP) problem (Schönemann, 1966), which finds a transformation matrix W^* with the following objective: $W^* = \min_W ||WA - B||$, s.t.: $W^TW = I$, where A and B are $c \times d$ embedding matrices. W^* is a $d \times d$ orthogonal matrix that transforms A through rotation and reflection, preserving the inner products of its rows. This objective minimizes the sum of the Euclidean distances between each pair of row vectors of A and B. This objective is applied only to the identified landmark set, L by finding: $W^* = \min_W ||WE_1^L - E_2^L||$, s.t.: $W^TW = I$, where E_i^L is the sub-matrix obtained by selecting the rows of E_i that correspond to landmark words.

3.4 Measuring Semantic Shift

Once the embeddings are aligned, senses can be examined for their rate of semantic shift, which is then used in downstream assessments like identifying lost and gained senses or detecting word-level shift. Using the landmarks and aligned embeddings, we use leave-one-out cross validation to select a cosine threshold that determines if a pair of vectors are semantically shifted (Algorithm Line 11). A token is stable (unshifted) from E_1 to E_2 if the cosine distance between the vectors for this token after alignment is below the threshold. In the base ConShift method, sense tokens that exist in both embeddings are compared for their shift rate. These correspond to senses already matched by clustering, which can be evaluated in more depth with alignment. Any sense present in a single embedding remains unmatched and is labeled loss or gained. With ConShift-M-J and ConShift-M-I, we evaluate all sense pairs of the form $(s_{1,i}, s_{2,j})$. Any stable pair is considered a match, allowing us to check for more nuanced cases that involve one sense mapping to multiple others like broadening and narrowing. As the two methods share the same alignment process, we can compare the impact of joint and individual clustering on shift detection. Figure 3 provides examples of match types across two corpora. ConShift only allows one to one mappings and hence cannot represent sense broadening, gain or loss. The full matching methods can identify all forms. We investigate their effectiveness through experiments in the next section. Table 1 provides a summary of the underlying differences

Method	Sense Assignment	Possible Sense Landmarks	Sense Shift Analysis
ConShift	Joint induction	Pairs from clustering	Only on paired senses
ConShift-M-J		All combinations of senses from E_1 and E_2	On all sense pairs
ConShift-M-I	Independent induction		

Table 1: A summary of the different algorithms introduced in this paper. We assess two forms of sense assignment through WSI, where senses are either shared or unique to the corpora. From there, the senses can be paired based on these clusters or matched at alignment. Valid matches are detected during shift analysis.

of the three algorithms we introduce in this paper.

For word level shift detection, we aggregate the variation level of individual senses to assign a graded score and binary label, modeling shift as a function of its senses. The graded score is computed as the weighted average of each sense pair's cosine distance, where the weight is proportional to the frequency of the given sense in the corpus. This includes all pairs, not just those that were matched. Binary labels are assigned based on whether the graded score is above the threshold learned from cross-validation.

3.5 Explaining Semantic Change

The most immediate application of ConShift is detection of semantic variation at the sense level, by mapping senses that are semantically related. When comparing corpora from two different time periods, absence of a sense in the later time period may indicate an obsolete meaning. Additionally, a sense from the past matching to multiple modern senses can indicate that the usage of that meaning has become broader, such as the word *broadcast* which initially meant *to scatter or sow seeds* and now has a related technological sense of *to send out or transmit*. If the two corpora are from different domains or communities, the absence and presence of senses can signal differences in language use. Cognates, for example, are words from different languages which share the same etymological origin, but have experienced different forms of semantic change. Thus, we assign individual labels to each sense based on the number of matches to detect sense loss, gain, broadening, narrowing and stable meanings (as shown in Figure 3).

Qualitatively, we highlight the forms of explanation possible at each step of our approach. Since our sense induction method is interpretable, cluster centroids from the lexical substitute vectors can be used to obtain the top substitute terms of each sense cluster. Additionally, the vectors closest to the centroid can be mapped to their original sentence form, to obtain exemplar usages of each sense. Once the corpora are transformed to type embeddings and aligned, we can visualize senses in the shared semantic space. For each sense, we can observe the meaning in the original context by its local neighbors, as well as the closest equivalent context of the other corpora through the mapped neighbors. If a concept is missing in the mapped corpus, the closest terms can give an estimation of how closely it's captured by the corpora. Given both the contextual and static neighbors, we construct a fuller picture of the senses and their relationships.

4 Diachronic Semantic Shift Evaluation and Explanation

In the following sections, we evaluate the effectiveness of ConShift and its variants on the task of diachronic semantic change detection using SemEval 2020 Task 1 (Schlechtweg et al., 2020) and LSCDiscovery in Spanish (Zamora-Reina et al., 2022). For the two joint clustered approaches, ConShift and ConShift-M-J, we assess the benefits of allowing additional matching between senses. Between the matched approaches, ConShift-M-J and ConShift-M-I, we investigate whether finding senses per corpus provides more accurate representations of senses and better matching.

4.1 SemEval Dataset

SemEval-2020 Task 1 is the first task on unsupervised lexical semantic change detection featuring an evaluation framework with gold standards (Schlechtweg et al., 2020). We use the English set, based on the Clean Corpus of Historical American English (CCOHA) (Alatrash et al., 2020), a preprocessed and lemmatized version of the Corpus of Historical American English (COHA). The corpus was split into two time periods: 1810–1860 and 1960–2010. 37 words were assigned binary and graded labels based on their amount of change between these time periods.

Word Change Detection Table 2 shows the task scores of our sense-aligned methods against the

Method	Binary	Graded
ConShift	0.703	0.482
ConShift-M-J	0.689	0.526
ConShift-M-I	0.730	0.608
Montariol et al. (2021)	-	0.456
Card (2023)	-	0.547
Tang et al. (2023)	0.730	0.589
Kutuzov (2020)	-	0.605
Rosin and Radinsky (2022)	-	0.627
Periti et al. (2024)	-	0.741
Cassotti et al. (2023)	-	0.757

Table 2: Word change detection results for SemEval-2020 Task 1, binary shift measured using accuracy and graded shift scored using Spearman correlation.

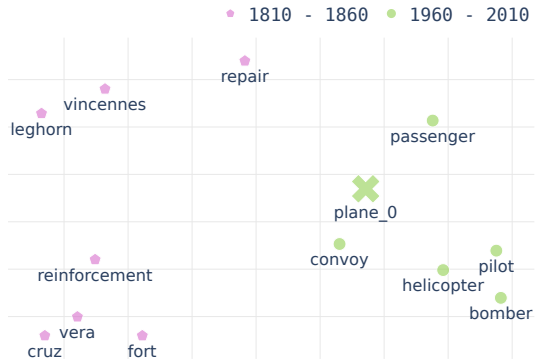

Figure 4: A contextual projection for a gained sense of the word *plane* from the SemEval dataset. Sense 0 from the 2000s, plane_0, and its local neighborhood is compared to the aligned neighbors from the 1800s.

current top performing approaches from the literature. The base ConShift approach has the lowest performance of our methods. For the two approaches with additional matching, the approach using senses from independent clustering, ConShift-M-I, performs better. The variation in performance can be explained by inspecting the difference in matching that occurs with each method. Most notably, despite having the ability to create additional matches, ConShift-M-J only creates single connections between senses for this dataset. We observe that joint clustering harms the modeling of certain senses once embedded when one corpus was underrepresented in the original sense cluster. This leads to a pair of senses where one is far less clear in its representation and occurs most often for words with gained meanings like *plane*, *record*, and *graft*. An example of this issue can be observed for the word *plane*, where 2 senses are found (airplane and a surface) when induced jointly. Table 3 provides a comparison of *plane*'s joint clustered senses, through the top lexical substitutes from WSI against the word embedding neighbors. The 1800s sense for airplane has an unclear neighborhood, leading it to not match to either of the 2000s senses. This is a false case of loss in the direction from 1800s to 2000s, when really the airplane sense has only been gained in the 2000s. When induced individually, we get clearer senses for each group. Two 1800s senses map to a more general 2000s sense, and the 2000s sense of an aircraft is identified as gained. Thus, improvement from ConShift to ConShift-M-J can be attributed to the change in sense matching that allows for different matches between senses.

The independent clustering approach, ConShift-M-I, produces many multi-way matches between senses, resulting in performance improvements over both joint methods. We conclude that the primary reason multi-way matches occur with ConShift-M-I and not ConShift-M-J is due to improved clustering that allows clusters to represent concepts without being influenced by the other corpus. Matching helps to distinguish senses that are truly shifted semantically, by considering the other possible polysemous senses that may be also be close in meaning.

Sense Change Evaluation The most common shift type identified for this dataset was sense gain. To inspect a single sense, we can project it after alignment into the shared semantic space of both embeddings using PCA. This allows us to observe the differences in concepts that each embedding captures. Figure 4 shows this for the target term *plane*, which gained a sense in the 2000s. Since the concept of an aircraft didn't exist in the 1800s, the closest contextual components are that of military implements and battle locations (Veracruz, Vincennes, and Leghorn). In addition, we observe cases of broadening, such as *record* shown in Figure 1. Record as a noun has a single etymological origin[2], with most senses connected to this meaning. We find three senses in the 2000s, each with their own unique context that still relates to the single 1800s sense of *record* as documented information. The SemEval dataset labels *record* as having gained a sense. Sense 2 from the 2000s, which covers a record of court proceedings, is matched to the 1800s sense with a cosine distance slightly

[2]https://www.oed.com/dictionary/record_n1

	Lexical Substitutes	1800s Embedding Neighbors	2000s Embedding Neighbors
0	line, direction, angle, shape, side, point	projection, horizontal, perpendicular, parallel, vertical, angle	particle, depth, surface, intricate, wavelength, curve
1	jet, flight, aircraft, pilot, ship, flying	whist, old-fashioned, smartly, loft, keziah, rubber	pilot, helicopter, passenger, airport, flight, aboard

Table 3: Comparison of terms representing senses of the term *plane* from joint clustering. MLM substitutes feature the terms with the highest average probability for that cluster, predicted from the BERT MLM task. Embedding terms are nearest neighbors.

Method	Binary Shift	Sense Gain	Sense Loss	Graded Shift
ConShift	0.651	0.00	0.610	0.275
ConShift-M-J	0.676	0.389	0.602	0.289
ConShift-M-I	0.709	0.431	0.630	0.449
Rachinskiy and Arefyev (2022)	0.716	0.511	0.688	0.735
Rombek (Not published)	0.687	0.520	0.681	0.535
Kudisov and Arefyev (2022)	0.658	0.520	0.610	0.209
Homskiy and Arefyev (2022)	0.655	0.591	0.582	0.676

Table 4: Results for LSCD Spanish subtasks, word shift level detection (Graded and Binary), and sense level detection (Gain and Loss). Graded change is reported as the Spearman correlation and the remainder are F1 score.

below the shift threshold. The remaining senses from the 2000s are considered shifted.

4.2 LSCDiscovery

We evaluate our sense-level labeling on the LSCDiscovery shared task for lexical semantic change detection in Spanish (Zamora-Reina et al., 2022). We use the evaluation dataset, which consisted of 60 human-annotated words that were assigned binary labels sense loss and gain, a binary change score and a graded change score.

Word and Sense Change Detection We report our results in Table 4 against the top performing methods from the task, ordered by their binary change score. Both the second and third best performing methods used the MLM approach for WSI with hierarchical agglomerative clustering, with different approaches for measuring shift (Rombek, Kudisov and Arefyev (2022)). Our best performing method, ConShift-M-I, has a high overall binary shift detection score, though it has a lower graded score, between the two HAC methods. The unmatched method, ConShift, performs unevenly in sense gain and loss labeling, due to the naive method for detecting sense changes based only on cluster membership. This method is too restrictive, so the approaches that match and consider the relationship between senses performs better.

5 Synchronic Sense Analysis

In addition to diachronic analysis, we provide an example of our method's ability to investigate differences in words from different communities. Since synchronic data has no inherent ordering, sense modeling looks for variations in usage, which can occur for a number of reasons, such as loan words gained from regional neighbors.

5.1 US - UK English Dataset

We look at dialectical differences between modern American and British English, based on Gruppi et al. (2021). 117 word pairs were identified for belonging to one of the following forms of word variation: 1) homonymy, 2) synonymy, and 3) alternate spellings (such as *color* and *colour*). Our UK corpus is the British National Corpus (BNC) XML Edition (of Oxford 2007) which contains a mix of news, fiction, and academic texts. The US corpus is the Corpus of Contemporary American English (COCA) (Davies, 2009) which contains text from newspapers, magazines, fiction, and academic texts. This dataset was not constructed based on a specific corpus. Rather, we use it to motivate exploration of possible synchronic patterns.

Evaluating Word Variation Using ConShift-M-I, we identify all three cases of variation in the dataset. For homonymy, we compare the US and UK senses of *saloon* in Figure 5. UK *saloon_1*

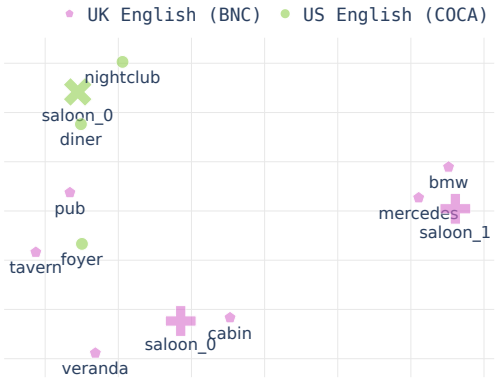

Figure 5: Senses of saloon from the US and UK. Saloon in the UK has two related senses of a high-class lounge (saloon_0) or sedan-style car (saloon_1), which originated from lounge-esque train cars. The US saloon_0 is a less reputable bar, so it is shifted from both UK senses.

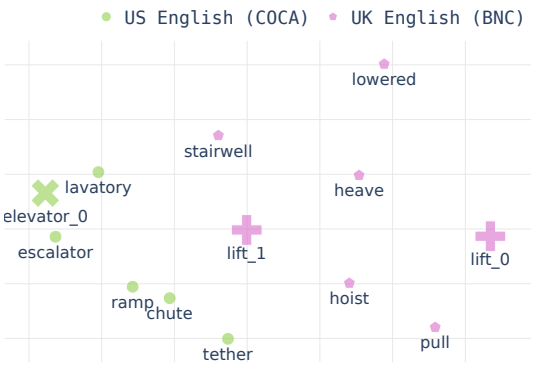

Figure 6: Projection of senses for *elevator* from the US corpus and *lift* from the UK corpus. We match noun senses, along with the related verb form (lift_0).

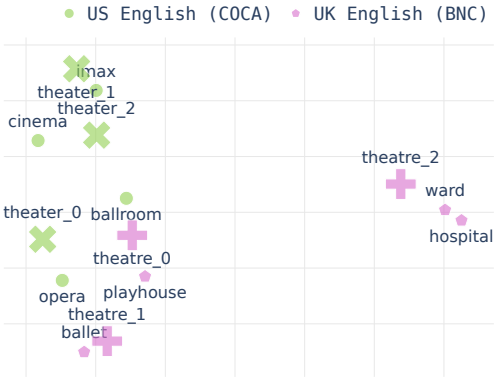

Figure 7: Projection of senses for US *theater* and UK *theatre*. Multiple senses match across corpora given the term similarity, with UK theatre_2 unmatched.

is the most shifted sense, which represents a style of car (known as a sedan in the US). Interestingly, the remaining senses are also unmatched. Though both senses refer to locations and share the same etymology, the UK *saloon_0* implies a more luxurious venue than US *saloon_0*. This is also the origin for UK *saloon_1*, as it derived from luxury train cars. Additional cases of homonymy that occur include *football* representing different sports and *casualty* having an additional meaning of an emergency room in the UK. Representing independent senses makes it easier to see the differences between usages that may seem related, as well as the relationship between meanings within a single corpus. For synonyms, we compare *elevator* from American English to *lift* in British English, shown in Figure 6. We find a common meaning in senses elevator_0 and lift_0 of a transportation method for moving between floors. These senses are detected as unshifted, along with the related meaning lift_1 as the action of raising something. Other pairs of this nature include *fall/autumn* and *gas/petrol*.

Figure 7 shows a case of comparing senses for a common word with different spellings, US *theater* and UK *theatre*. Most senses are related to the concept of a room where performances happen and grouped together. The UK has a unique but related sense of a room where operations occur. In some instances, we don't observe the variation that is stated in the dataset for a given target. This is generally because the unique sense is not represented in its intended corpus or has occurred in both corpora. For example, the target *hamper* contains the sense of a wicker basket that holds laundry in the US and one that holds food in the UK, though the latter sense is missing. Most often this issue arises from missing UK senses, like the synonym pair of *mail* and *post*. Conversely, the US corpora sometimes contains UK usages as well, as is the case with *trolley*, a British word for shopping cart. While these issues limit the current quantitative ability of the dataset, they highlight the nuances that come with studying synchronic variation in a pair of corpora.

6 Conclusions

We introduced ConShift, a set of methods for modeling sense-level semantic change across corpora in synchronic and diachronic settings. ConShift provides explanations of semantic change through sense comparisons and conceptual mapping. We plan to further assess our alignment method using alternate sense labeling methods and shift detection measures, and investigate more niche cases of sense variation.

Limitations

We make a number of assumptions in our method and the applicability of the method will depend on whether these assumptions are true or not. Our first assumption is that type of change we model when performing perturbations in the self-supervised alignment is sufficient for capturing the various ways senses may be shifted. The second assumption involves the sense induction models providing an accurate representation of senses. Even though our model is agnostic to the method used for sense induction and we allow a single sense to be mapped to zero to multiple senses in the other corpus, we are still limited by the quality of the initial clusters we obtain. Once senses are induced, they are transformed to types and not decomposed any further. The underlying contextual embedding may not be well suited to the given corpus and may need to be retrained on it to be more effective. However, the impact of such adjustments are hard to measure due to lack of English language datasets with sense shift labels. While datasets with word-level semantic change labels exist, these are also rather small and do not necessarily capture many words with multiple evolving senses. Our method requires the selection of a number of words in advance for sense level analysis, as using MLM for the whole corpus would be too expensive computationally. Additionally, many word embedding models filter out words that occur infrequently in a corpus. Breaking a word into multiple senses will make each less frequent and may put its count below the threshold. This requires balancing goals between the embedding and induction step.

Ethics Statement

Language use differs between groups separated by time, cultural background, mode of communication and topics of interest among many others. Hence the study of semantic variation must take into account whether the changes captured by a specific method can be classified as a true semantic shift, i.e. a use of the word for a different concept. Even lexicographers may disagree on this question sometimes, hence this is a truly difficult problem. Additionally, alternative explanations may be available for the difference captured by our algorithm. For example, different datasets may not be easily comparable due to differences in the breadth of topics, the communication modes covered or the time periods. This is especially important in synchronic settings and arbitrary datasets.

Given the ambiguity of language and insufficiency of labels especially for word level change, semantic variation detection methods are best applied for easing analysis of datasets, not by automated decision making. Our methods can be very useful for comparing and understand corpora from different groups, analyzing limitations of large language models and applications based on them when faced with such semantic differences in word usage. However, it should never be used to make a value judgment on whether any specific usage is proper or improper. Our aim should be to develop tools to help understand each other better, not to narrow down which usages are appropriate to incorporate into models.

References

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte Im Walde. 2020. CCOHA: Clean Corpus of Historical American English. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6958–6966.

Asaf Amrami and Yoav Goldberg. 2019. Towards better substitution-based word sense induction. ArXiv:1905.12598 [cs].

Nikolay Arefyev and Vasily Zhikov. 2020. BOS at SemEval-2020 Task 1: Word Sense Induction via Lexical Substitution for Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 171–179, Barcelona (online). International Committee for Computational Linguistics.

Robert Bamler and Stephan Mandt. 2017. Dynamic Word Embeddings. In *Proceedings of the 34th International Conference on Machine Learning*, pages 380–389. PMLR. ISSN: 2640-3498.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 Diachronic Lexical Semantics (DIACR-Ita) Task. In Valerio Basile, Danilo Croce, Maria Maro, and Lucia C. Passaro, editors, *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020*, pages 411–419. Accademia University Press.

José Canete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:1–10.

Dallas Card. 2023. Substitution-based semantic change detection using contextual embeddings. In *Proceedings of the 61st Annual Meeting of the Association*

for Computational Linguistics (Volume 2: Short Papers), pages 590–602, Toronto, Canada. Association for Computational Linguistics.

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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. *Advances in Neural Information Processing Systems*, 32.

Mark Davies. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190. Publisher: John Benjamins.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Publisher: Association for Computational Linguistics.

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

Maurício Gruppi. 2022. *A Computational Approach to Lexical Semantic Shift Across Time and Domain: Methods and Applications*. Ph.D., Rensselaer Polytechnic Institute, United States – New York. ISBN: 9798368448565.

Maurício Gruppi, Pin-Yu Chen, and Sibel Adali. 2021. Fake it Till You Make it: Self-Supervised Semantic Shifts for Monolingual Word Embedding Tasks. *AAAI*, 35(14):12893–12901.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, pages 67–71.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501.

Daniil Homskiy and Nikolay Arefyev. 2022. DeepMistake at LSCDiscovery: Can a Multilingual Word-in-Context Model Replace Human Annotators? In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 173–179, Dublin, Ireland. Association for Computational Linguistics.

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

Vani Kanjirangat, Sandra Mitrovic, Alessandro Antonucci, and Fabio Rinaldi. 2020. SST-BERT at SemEval-2020 Task 1: Semantic Shift Tracing by Clustering in BERT-based Embedding Spaces. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 214–221, Barcelona (online). International Committee for Computational Linguistics.

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.

Artem Kudisov and Nikolay Arefyev. 2022. BOS at LSCDiscovery: Lexical Substitution for Interpretable Lexical Semantic Change Detection. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 165–172, Dublin, Ireland. 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.

Andrey Kutuzov. 2020. *Distributional word embeddings in modeling diachronic semantic change*. Doctoral thesis. Accepted: 2020-11-16T12:34:15Z.

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.

Noa Yehezkel Lubin, Jacob Goldberger, and Yoav Goldberg. 2019. Aligning Vector-spaces with Noisy Supervised Lexicons. 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 460–465.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, 26.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1020–1029.

Stefano Montanelli and Francesco Periti. 2023. A Survey on Contextualised Semantic Shift Detection. *arXiv preprint arXiv:2304.01666*.

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.

Francesco Periti, Pierluigi Cassotti, Haim Dubossarsky, and Nina Tahmasebi. 2024. Analyzing semantic change through lexical replacements.

Francesco Periti, Alfio Ferrara, Stefano Montanelli, and Martin Ruskov. 2022. What is Done is Done: an Incremental Approach to Semantic Shift Detection. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 33–43.

Martin Pömsl and Roman Lyapin. 2020. CIRCE at SemEval-2020 Task 1: Ensembling Context-Free and Context-Dependent Word Representations. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 180–186, Barcelona (online). International Committee for Computational Linguistics.

Maxim Rachinskiy and Nikolay Arefyev. 2022. GlossReader at LSCDiscovery: Train to Select a Proper Gloss in English – Discover Lexical Semantic Change in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 198–203, Dublin, Ireland. Association for Computational Linguistics.

Guy D. Rosin and Kira Radinsky. 2022. Temporal attention for language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1498–1508, Seattle, United States. 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, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091.

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

Justin Solomon. 2018. Optimal transport on discrete domains. *AMS Short Course on Discrete Differential Geometry*. Publisher: Notices of the AMS.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. *Computational approaches to semantic change*, 6(1). Publisher: Language Science Press Berlin.

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.

Zi Yin, Vin Sachidananda, and Balaji Prabhakar. 2018. The Global Anchor Method for Quantifying Linguistic Shifts and Domain Adaptation. *Advances in Neural Information Processing Systems*, 31.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 149–164, Dublin, Ireland. Association for Computational Linguistics.

Jinan Zhou and Jiaxin Li. 2020. TemporalTeller at SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection with Temporal Referencing. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 222–231, Barcelona (online). International Committee for Computational Linguistics.

A Appendix

A.1 Experiment Details

In this section, we report the parameters that were selected at each stage of our pipeline and the reasoning behind them.

Assignment Method	Alignment Method	Accuracy (Binary)	Spearman (Graded)
MLM+HAC	ConShift	0.703	0.482
	ConShift-M-J	0.689	0.526
	ConShift-M-I	0.730	0.608
BERT Embed + k-means	ConShift	0.770	0.522
	ConShift-M-J	0.716	0.545
	ConShift-M-I	0.716	0.512

Table 5: Comparison of sense assignment approaches on word change detection performance for SemEval-2020 Task 1.

Sense Assignment The first of our approaches is the MLM-based WSI that uses hierarchical agglomerative clustering (HAC). Amrami and Goldberg (2019) sample the probability distributions to get a bag of words of common substitutes prior to clustering. We work directly with the probabilities output by the model. Since most terms in the distribution will have a probability near 0 for a given masked token, we reduce the vector size by subsetting the vocabulary (Arefyev and Zhikov, 2020; Card, 2023). First, stopwords and non-alphabetical terms like symbols and numbers are removed. Secondly, we remove the terms that are predicted at very high and low frequencies. For all prediction vectors of a target, we find the rate that a term occurs with a probability above 0.01%. Any term in less than 3% or more than 90% of prediction vectors is removed (Arefyev and Zhikov, 2020; Kudisov and Arefyev, 2022). This eliminates infrequent words with low probabilities that are unlikely to impact clustering, as well as highly frequent words that may over-influence cluster merging. Following this process, we subset the vectors using the reduced vocabulary and normalize them. This procedure is done for each target separately, so the final length of vectors vary.

Once all lexical substitute vectors are obtained for a target, hierarchical agglomerative clustering is used to create an initial set of sense clusters by flattening the model. We start with 15 clusters to replicate the approach in Amrami and Goldberg (2019), then iteratively merge clusters based on the centroid distance. We calculate cluster distances with average linkage and use cosine similarity to measure the distance between normalized lexical substitute vectors. Sense cluster size is measured by the percentage of vectors within the cluster from the overall set, where any cluster below $n\%$ will be merged. To find the best value of n for each target, we search over a full range of percentages until the vectors are clustered into a single sense (often around $n = 20$) and evaluate using the silhouette score. If the silhouette score was never above 0, one sense is created for that target. A hard lower bound of at least 5 vectors is set to avoid singleton clusters and match the minimum count parameter at the word embedding step.

The second approach we evaluate is BERT embedding with k-means clustering (Giulianelli et al., 2020). For each term, we create k clusters ranging from 2 to 11. Both methods determine the best clustering through silhouette score, using euclidean distance for its metric. For English tasks, we use the bert-base-uncased model with 12 layers, 768 hidden dimensions and 110M parameters (Devlin et al., 2018). For Spanish, we use BETO (Canete et al., 2020), a BERT model trained with the whole word masking technique on a large Spanish Corpus (3B words). This model has the same dimensions as bert-base, with a similarly sized vocabulary. The MLM+HAC method was performed using the Scipy implementation and k-means clustering used scikit-learn.

We compare results of the two sense assignment methods for SemEval-2020 Task 1 (4.1) in Table 5. Given the small size of the SemEval dataset, the scores for binary change detection have little variation across the combination of approaches. For graded detection, independent clustering performs best for both methods of WSI. For jointly clustered senses, performance differs between the methods of aligning with and without matching.

Word Embedding We use Gensim's implementation of Word2Vec [3] to create word embeddings. For all datasets, we use a vector size of 300 and window of 10. Selection of the minimum frequency is influenced by the minimum sense size, as discussed in Section 3.2. For the SemEval and LSCDiscovery tasks, this value is set to 5. TFor the larger corpora

[3] https://radimrehurek.com/gensim/models/word2vec.html

Dataset Name	# Pos.	# Neg.	Shift Rate
SemEval	100	25	0.1
LSCDiscovery	50	50	1
US-UK	100	100	0.1

Table 6: Selected alignment learning parameters (number of positive samples, number of negative samples and rate of artificial shift) for each dataset.

from the US-UK dataset, we set the minimum frequency to 50.

Alignment For the iterative learning on pseudo-shifted vectors, described in Section 3.3, we select the number of positive and negative samples used for training, as well as the rate of artificial shift that is applied to a selected vector. We performed a parameter sweep to select values for our final reporting, based on the reported ranges from Gruppi (2022). For positive samples that are created using artificial shift, we select from [50, 100]. Negative samples, which draw from the landmarks, were from the range of [25, 50, 100]. For the shift rate, we check [.1, .5, 1]. We report the final parameters that we use for each task in Table 6. When determining the best cosine threshold through leave-one-out cross validation, we search from 0 to 1 with a step size of .05.

Algorithm 1 : Pseudo-code of ConShift-M where two corpus embeddings, $E1$ and $E2$, are aligned after selecting landmarks from their shared vocabulary, V and all sense pairs. Senses may be shared (ConShift-M-J) or independent (ConShift-M-I), as they are paired through a Cartesian product. $Pairs$ consists of all word tuples between $E1$ and $E2$, where $pair = (word_{1,a}, word_{2,b})$, and a and b are two matching words or paired senses. In the base approach of ConShift, we skip the construction of additional sense pairs from lines 7 and 8 and use the original V_{shared}, containing only words and sense tokens that are the same in both embeddings. The shift of a set of target terms T is detected after alignment.

1: **Data:** $E1, E2, V1, V2$
2: **Result:** Shift predictions, P, and sense mappings, M_s

Selecting Landmarks

3: $V_{shared} \leftarrow (V1 \cap V2)$
4: Pairs$_{shared} \leftarrow [(w, w)] \forall w \in V_{shared}$
5: $S1 \leftarrow$ find_senses$(V1)$
6: $S2 \leftarrow$ find_senses$(V2)$
7: Pairs$_{sense} \leftarrow S1 \times S2$
8: Pairs $\leftarrow$ Pairs$_{shared} \cup$ Pairs$_{sense}$
9: $L \leftarrow$ landmark_search$(E1, E2,$ Pairs$)$ ▷ Described in Section 3.3.1

Alignment

10: $E1, E2' \leftarrow$ orthogonal_Procrustes$(E1, E2, L)$ ▷ Described in Section 3.3.2

Measuring Semantic Shift

11: cosine-threshold $\leftarrow$ cross-validation$(E1, E2', L)$
12: Shifted$_s \leftarrow [p] \forall p \in$ Pairs$_{sense}$ **if** $cos\_dist(p) >$ cosine-threshold
13: Unshifted$_s \leftarrow [p] \forall p \in$ Pairs$_{sense}$ **if** $cos\_dist(p) <$ cosine-threshold
14: $M_s \leftarrow$ Shifted$_s \cup$ Unshifted$_s$
15: $P \leftarrow$ weighted_average(M_s)
16: **return** P, M_s