Current Semantic-change Quantification Methods Struggle with Discovery in the Wild

Khonzoda Umarova **Lillian Lee** **Laerdon Kim**
Dept. of Computer Science, Cornell University
ku47@cornell.edu llee@cs.cornell.edu lyk25@cornell.edu

Abstract

Methods for lexical semantic-change detection quantify changes in the meaning of words over time. Prior methods have excelled on established benchmarks consisting of pre-selected *target* words, chosen ahead of time due to the prohibitive cost of manually annotating all words. However, performance measured on small curated wordsets cannot reveal how well these methods perform at *discovering* semantic changes among the *full* corpus vocabulary, which is the actual end goal for many applications.

In this paper, we implement a top-k setup to evaluate semantic-change *discovery* despite lacking complete annotations. (At the same time, we also extend the annotations in the commonly used LiverpoolFC and SemEval-EN benchmarks by 85% and 90%, respectively). We deploy our evaluation setup on a battery of semantic-change detection methods under multiple variations.

We find that when presented with a *natural* distribution of instances, *all* the methods struggle at ranking known large changes higher than other words in the vocabulary. Furthermore, we manually verify that the majority of words with high detected-change scores in LiverpoolFC do not actually experience meaning changes. In fact, for most of the methods, less than a half of the highest-ranked changes were determined to have changed in meaning. Given the large performance discrepancies between existing-benchmark results and discovery "in the wild", we recommend that researchers direct more attention to semantic-change discovery and include it in their suite of evaluations. Our annotations and code for running evaluations are available at https://github.com/khonzoda/semantic-change-discovery-emnlp2025.

1 Introduction

Semantic change is the phenomenon of change in meaning of words over time. In recent years, especially in the light of the emergence of modern contextualized representations for language and meaning, there has been revitalized interest in lexical semantic-change detection methods (Periti and Montanelli, 2024). These typically aim to identify if or by how much words have shifted meaning between two or more consecutive time periods (Schlechtweg et al., 2020).

Prior work has established benchmarks for evaluating methods that detect lexical semantic change (Del Tredici et al., 2019; Schlechtweg et al., 2020). Most commonly, a benchmark consists of sub-corpora from different time periods and a set of *target* words, T, manually annotated for presence/absence/level of semantic change. Performance is measured by the degree to which a method's semantic-change judgments align with human judgments for words in T. With recent advances in language modeling, researchers have developed methods that excel at these benchmarks (Periti and Tahmasebi, 2024).

However, performance measured on a small curated set of words doesn't necessarily predict how well these methods would perform "in the wild." Instead, in this paper, we consider **semantic-change *discovery* (SCDisc)**. In this setting, researchers would apply discovery methods to organically identify and extract potential meaning changes from corpora, perhaps passing the detected words on for further inspection by domain experts such as cultural historians or literary theorists employing distant reading (Kim et al., 2014; Hamilton et al., 2016a).

Evaluating semantic-change discovery is quite challenging because, to our knowledge, there are no existing non-curated corpora in which all or a sufficiently large sample of words have been systematically annotated for degree of change. Our proposal does make use of the annotated vocabulary subsets attached to existing benchmarks: we first measure how highly the words in T that were

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 35354–35367
November 4-9, 2025

manually annotated as having changed the most — call these $T^* \subset T$ — are ranked in relation to the rest of the full corpus vocabulary, V. Across multiple methods, this evaluation shows that a large number of words from $V \setminus T$ outrank the words in T^*. Moreover, in one dataset, for the majority of the methods, *none* of the 15 top-scoring words are from T^*; in another dataset, across *all* methods at most two T^* words appear in the top-scoring 100.

It seems unlikely that all or most of the highly-ranked non-target words in fact underwent greater change than the annotated words (i.e., that these are all true but un-annotated positives). Nonetheless, we address this counter-explanation for the poor results of existing methods by further undertaking the task of manually annotating the top-k results of a suite of methods on a frequently-used benchmark (Del Tredici et al., 2019). By suitably adapting the labeling protocol of Del Tredici et al. (2019), we obtain human annotations where (a) judgments about semantic change are done in consideration of a group of sentences from each time period, as opposed to pairwise judgments as in the Schlechtweg et al. (2020) annotation scheme; and (b) annotator time is used more efficiently when no semantic change occurs.

The prior work closest to ours, Kurtyigit et al. (2021), applies top-k evaluation to two discovery methods on the SemEval2020 German corpus (Schlechtweg et al., 2020), wherein the relative distribution of changed vs. (assumed to) not-have-changed words is controlled. Since subsequent contextualized-representation-based methods achieve state-of-the-art performance on lexical semantic-change benchmarks (Periti and Tahmasebi, 2024), we extend our evaluation to a broader range of recent semantic-change detection methods. In our investigation, we work with two corpora. SemEval-EN (Schlechtweg et al., 2020) is in English, but, like the German SemEval corpus mentioned above, is curated to sample hypothesized changes while balancing them with background terms. LiverpoolFC (Del Tredici et al., 2019), in contrast, is unfiltered, consisting of all comments and submissions on a sub-reddit of the same name that fall within specified time periods.

Contrary to SCDisc performance in SemEval-EN, where over half of the top-ranked terms were verified via annotation to have in fact undergone semantic change, in LiverpoolFC these methods often struggle at *discovering* true semantic changes: less than half and in many instances less than a third of the top-ranked words in LiverpoolFC are genuine semantic changes.

Contributions summary **[i]** First, we propose a top-k evaluation paradigm and apply it to quantify the *discovery* performance of a broad range of (contextualized language model $\times$ lexical semantic-change detection method $\times$ secondary adjustment/filtering technique) combinations depicted in Figure 1. **[ii]** We find that the performance of these methods varies significantly by evaluation criterion[1] and dataset. When it comes to ranking all already-known high changes (i.e., T^*) well, semantic-change detection methods flounder in both datasets, more so in SemEval-EN than in LiverpoolFC. On the other hand, when it comes to finding new un-annotated semantic-change instances, they do well in SemEval-EN but struggle on LiverpoolFC. **[iii]** As a by-product, we extend both SemEval-EN and LiverpoolFC with additional annotations created with a somewhat novel annotation procedure.

2 Preliminary Notation

Existing benchmarks (Schlechtweg et al., 2018; Del Tredici et al., 2019; Schlechtweg et al., 2020; Kutuzov and Pivovarova, 2021) consist of (i) a corpus C, over some vocabulary V, that is divided into temporally distinct sub-corpora, $C = C_1, C_2, \ldots$ and (ii) a (relatively small) vocabulary subset $T \subset V$ where for each $t \in T$, a label $\ell(t; C) \in [0, 1]$ has been derived by a team of human annotators. Often, T is constructed initially from words hypothesized to have changed in meaning, and then further balanced with a set of control words (i.e., distractors). The performance of a semantic-change scorer $\hat{f}$, which is a function of any vocabulary item $v \in V$ and the temporally-divided corpus C, is evaluated by the difference between $\hat{f}(t; C)$ and $\ell(t; C)$ for $t \in T$.[2] However, such evaluation does not illustrate how semantic change methods perform on words outside of T.

3 Related Work

With the goal of simulating "discovery"-like evaluation, Zamora-Reina et al. (2022) introduced a SCDisc benchmark where methods need to rank

[1] See more details about our evaluation criteria in §4.1 and §4.2

[2] For brevity's save, we will typically omit explicit indication of the dependence of $\hat{f}(\cdot)$ and $\ell(\cdot)$ on C.

changes for a large slice of V that has the annotated set T hidden within it. However, the final evaluation is still based on the curated set T.

Some researchers do inspect and validate selected words with the best $\hat{f}$ scores, rather than only focus on T (Kim et al., 2014; Hamilton et al., 2016b; Kutuzov et al., 2022). Kutuzov et al. (2022) find that the highest-$\hat{f}$ words according to contextualized-embeddings-based methods often still contain words that, according to post-hoc inspection, did not actually exhibit meaning change. They identify as culprits such phenomena as words that experience high variance in usage context, a "data burst" of one specific usage context in one time bin, and syntactic changes.

Kurtyigit et al. (2021): Semantic-Change Discovery

In an approach similar to ours, Kurtyigit et al. (2021) evaluate the discovery capabilities of methods based on static (SGNS (Mikolov et al., 2013)) and contextualized (BERT (Devlin et al., 2019)) embeddings. From there, they select the 30 highest-$\hat{f}$ words from each of the two methods (at the best parameter settings) and annotate them similarly to how Schlechtweg et al. (2020) do. The annotations reveal that 67% and 57% of the discovered words indeed undergo semantic change. Building upon Kurtyigit et al. (2021), we consider a variety of contextualized language models beyond BERT, such as XLM-R (Conneau et al., 2020) and XL-LEXEME (Cassotti et al., 2023). We further expand our evaluations to a broader range of contextualized-representation-based methods. In addition, we complement these methods with secondary techniques (such as permutation tests and scaled change metrics) that aid semantic-change detection to see if these techniques also aid SCDisc.

In our approach we evaluate SCDisc on two (quite distinct) datasets, not just one. The first, SemEval-EN (Schlechtweg et al., 2020), is a benchmark similar to that used by Kurtyigit et al. (2021). The other is LiverpoolFC (Del Tredici et al., 2019), where the subcorpora consist of all social-media posts and comments (i.e., no filtering) for a given subreddit during certain time periods. While Kurtyigit et al. (2021) discover new changes among a subset of 500 words from the vocabulary, we consider a much larger subset of the vocabulary to better capture the difficulty of the SCDisc task.

The final distinction of our approach from that of Kurtyigit et al. (2021) lies in how we gather annotations that would be used to verified discovered changes. Kurtyigit et al. (2021) follow the annotation procedure introduced by Schlechtweg et al. (2020) as part of SemEval-2020 Task 1. In contrast, we adapt a variant of the procedure by Del Tredici et al. (2019), which we discuss in more detail in §4.2.

4 Evaluation Metrics

In this section, we describe our SCDisc evaluation framework. Recall that the primary obstacle to measuring either the precision or recall of an SCDisc method $\hat{f}$ is that obtaining gold-standard labels for the entire vocabulary with respect to the collection of possibly large benchmark sub-corpora is infeasible; what is available is only a (relatively small) vocabulary subset $T \subset V$ where for each $t \in T$, a label $\ell(t) \in [0, 1]$ has been derived by a team of human annotators. We start by ranking words in T^* relative to those in $V \setminus T^*$ (§4.1). Next, we verify whether the top-k highest-$\hat{f}$-scoring items are in fact true instances of semantic change or not via post-hoc human annotation to fill in missing labels. (§4.2).

4.1 Ranking Changes for Discovery

In this subsection, we give the details of our ranking approach; the next subsection addresses the potential pitfalls of using ranking alone to approximate SCDisc performance.

Existing benchmarks provide an annotated set T. It is natural to highlight within T the subset with the largest human-label values: $T^* = \{t \in T \mid \ell(t) > \beta\}$ for threshold parameter β. For example, one might choose $\beta = 0$, making T^* precisely the set of words in T that exhibit any positive degree of semantic change; alternately, one could require greater annotator "confidence" by setting a higher threshold.[3] Previous work then evaluates the ranking induced by a scorer on T, essentially checking whether the $\hat{f}$-induced positions of the words in T^* are higher (better) than the positions of the terms in $T \setminus T^*$. We instead, as described in the Introduction, consider measuring how well a method $\hat{f}$ ranks the words in T^* against all of $V \setminus T^*$.

An important complication to note is the role of frequency effects. We observed, as have others (Dubossarsky et al., 2017; Noble et al., 2021; Kurtyigit et al., 2021; Card, 2023), that words of very

[3] In our experiments, we choose β values that result in $|T^*| = 15$: 0.59 for LiverpoolFC, 0.4 for SemEval-EN.

high and/or very low frequencies are particularly difficult for semantic-change detection methods:

1. A low-frequency word, by definition, occurs rarely, so there is less opportunity in a given time period to observe it in all its possible contexts. Consequently, in a later time period, the word may appear in new contexts — thus appearing to have changed semantically — even though the contexts in the second time period are novel only because we didn't have enough samples from the earlier time period.[4]

2. Counter-intuitively, high-frequency words such as function words also exhibit a type of confounding variance. We don't think of function words as semantically changing. But, because they appear in a huge variety of contexts across different time periods, they often receive high $\hat{f}$ scores, as shown by Noble et al. (2021) .

For these reasons, we follow peer researchers in explicitly excluding high- and low-frequency words from V. (Kurtyigit et al., 2021; Card, 2023, inter alia).[5]

We run each semantic-change scorer $\hat{f}$ on the frequency-filtered version of V plus the terms in T^*, and sort the terms into list $W_{\hat{f}}$ by their semantic-change scores, most-changed first. We then compute $\hat{f}$'s *discovery rate@*r as the recall of T^* among the top-r words of $W_{\hat{f}}$:

$$\phi_{\hat{f}}(r) = \frac{|\{w \in W_{\hat{f}}[:r]\} \cap T^*|}{n}. \qquad (1)$$

(For brevity's sake, we henceforth omit explicitly indicating the dependence on $\hat{f}$.)

As $r \leq |W|$ increases, the value of $\phi(r)$ approaches 1. We can say that a method performs poorly at discovery if $\phi(r)$ lies significantly below the $y = \frac{r}{n}$ line . . . *under the assumption* that all or most of the words in $W[:r] \setminus T^*$ are not instances of genuine semantic change. But does this assumption actually hold? The next subsection addresses this question.

4.2 Human Annotation of the Top-k Changes

In practice, in the absence of annotation of the full V, we don't know whether words in $V \setminus T$ exhibit semantic change or not. Hence, we complement $\phi(r)$ with an alternative metric, $\psi(k)$, that quantifies SCDisc performance as the percentage of the top-k highest-$\hat{f}$-scoring words that are verified by (potentially post-hoc) human annotation to have actually semantically changed.

Annotation based on group majority sense For the annotation, we start with the setup of Del Tredici et al. (2019): for a given word (type), annotators are presented with two groups, one from each time period (sub-corpus), of 5 sentences each (or as many as possible in the case that fewer than 5 are available), all of which contain the given type.

Our procedure differs from that of Del Tredici et al. (2019) in the following two respects. First, to cover a larger part of the corpus, the sentence groups for the two periods are resampled for each annotator. Second, instead of asking whether the meaning of the given word in group 1 is different from the meaning in group 2, we break the question up into several parts (details in Appendix A):

(a) Does group 1 have a majority sense? Group 2?

(b) Is the majority sense in group 1 different from that in group 2?

(c) Are you confident about the difference or lack thereof in the majority senses?

(d) What are the sentences whose senses appear in group 1 but not in group 2 (and vice versa)?

We aggregate the semantic change score per word, w, across all annotators by computing the average of labels for (b), and *extend* the label function by defining $\ell(w)$ as this average. In cases where there isn't a majority sense in either group 1 or group 2, we automatically re-assign the label for (b) to be 1 if there is at least one sentence listed under (d) and 0 otherwise.

This annotation setup is different from that of SemEval (Schlechtweg et al., 2020), where annotations are provided for individual pairs of use cases of the given word. We choose the Del Tredici et al.

[4] A concrete example: Suppose word w has two equally likely senses and appears only once in sub-corpus C_1 and once in C_2. Then there is a 50% likelihood that w has a different sense in the two time periods. On the other hand, suppose word w' likewise has two equiprobable senses s_1 and s_2, but appears 1000 times in each of the periods. Then the likelihood that w' has only s_1 in one period and only s_2 in the other is minimal.

[5] For the LiverpoolFC dataset, we only allow words with frequencies in the range [12, 2700]; for SemEval-EN the corresponding range is [20, 14000]. These intervals represent $\left[\frac{1}{2} \min \mathrm{fr}_T, 2 \max \mathrm{fr}_T\right]$, where $\min \mathrm{fr}_T$ and $\max \mathrm{fr}_T$ are the smallest and largest word frequencies in T, respectively.

(2019) setup for two reasons. First, annotating semantic change via group-based comparisons is a more efficient use of annotator time. Second, after conducting pilot runs of both the SemEval and Del Tredici et al. (2019)'s annotation schemes, we found that group vs. group allowed our annotators to better capture semantic change *relative* to the range of meanings of the word within time periods: by comparing groups of sentences, the annotators can decide if the variation in meaning is due to a different sense being introduced or due to general non-diachronic polysemy of the word. In contrast, sentence vs. sentence requires the annotators to make a large number of pairwise comparisons, and due to the linear manner of viewing new contexts, the annotators would not get the same perception of *holistic* change.

Special case: proper names We looked at the top-ranked terms produced by semantic-change detection methods and found that many of them correspond to names/nicknames for people, brands, and companies. The context changes for such words are often due to real-world events or changes in situations surrounding the person/entity associated with the word.

Consider the example of the former football player *Fernando Torres*, who belonged to the Liverpool Football Club between 2007-2011 before leaving to join Chelsea Football Club and then Atlético Madrid in 2016. Further, different events surrounding Torres' career also influence the usage context of the term "Torres". Yet, the sense, or grounding, of "Torres" remains unchanged, namely, the person Fernando Torres. Hence, like Kurtyigit et al. (2021), we exclude proper names from the annotations list and computation of $\psi(k)$.

Annotations-based discovery performance After filtering out proper names from $\hat{f}$-sorted list[6] W, we consider the top-k remaining terms: $W'[:k]$. In our experiments, we select k=15 as a reasonable number of top-ranked changes in a scenario where a researcher uses top-ranked changes as candidates for further investigation. As before, we extend the label function $\ell(\cdot)$ to $W'[:k]$ by using average human-annotation scores. We then binarize $\ell(w)$ to indicate whether a word changed in meaning or not, using a threshold parameter θ. To set the value of θ, we additionally gather annotations for the 5 most changed and 5 least changed words in T, and set θ to the value that best separates these two groups. Altogether, our annotations-based metric for SCDisc becomes

$$\psi(k) = \frac{|\{w \in W'[:k] \mid \ell(w) \geq \theta\}|}{k} \quad (2)$$

(dependencies on $\hat{f}$ omitted for brevity).

Annotators We recruit English speakers as annotators from a pool of volunteers that includes individuals with a university-level education as well as university students interested in research. Eight annotators followed our annotation procedure to label 76 words (types) from the SemEval-EN dataset, such that each type has at least three annotations. Similarly, ten annotators provide annotations for 83 types from the LiverpoolFC dataset.

Our annotations are available at https://github.com/khonzoda/semantic-change-discovery-emnlp2025.

5 Experimental Setup

5.1 Data

SemEval-EN is the largest English-language benchmark dataset for semantic-change detection (Schlechtweg et al., 2020). Its two subcorpora are from the 19th century (6.5M tokens) and the 20th century (6.7M tokens). For the 37 target lemmas provided, we select graded semantic-change annotations. We extract 82 distinct word types associated with these 37 lemmas and choose the highest-$\ell(\cdot)$-valued 15 to be our T^*.

LiverpoolFC is a dataset constructed by Del Tredici et al. (2019) from submissions and comments from the r/Liverpool subreddit, separated into two periods spanning 2011-2013 (8.5M tokens) and 2017 (11.9M tokens). The dataset provides graded semantic-change values from human annotations for $|T| = 97$ words. From these we set T^* to be the $n = 15$ types with highest $\ell(\cdot)$ values.

5.2 Models

We select four models commonly used in prior work: BERT and mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and XL-LEXEME (Cassotti et al., 2023)[7]. These were also the basis for Periti et al.'s (2024) systematic comparison of

[6] Based on preliminary experiments we select BERT and one method from each representation type to get lists for annotations; details about methods are in §5.3.

[7] With the exception of XL-LEXEME, we fine-tune models for 5 epochs on each dataset.

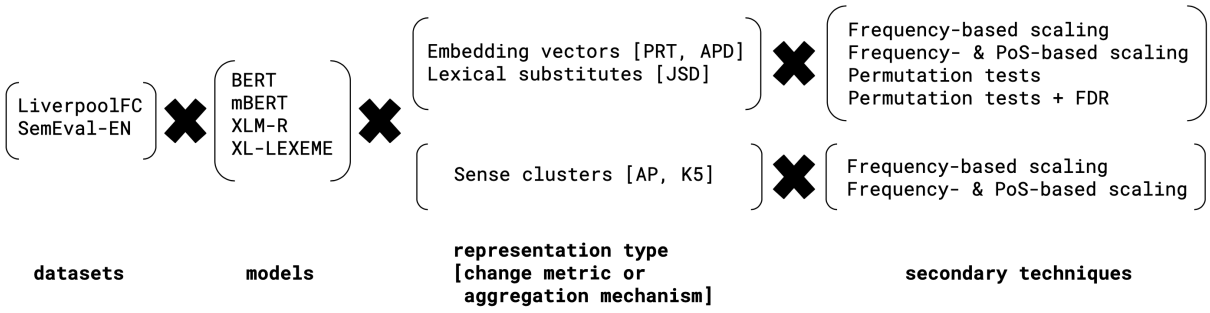


Figure 1: We conducted a wide range of experiments spanning various combinations of token/lemma representation types, change metrics, scaling, statistical significance tests, models and datasets.

semantic-change methods based on contextualized representations.

5.3 Methods

Next, we outline our selection of methods for graded semantic change detection from three categories: embedding-, substitutes-, and sense clusters-based representations. We describe each method in terms of how word usage instances are *represented*; how a word's representations within a time period are *aggregated*; and finally, how *assessment metrics* quantify semantic change between time periods (Periti and Montanelli, 2024).

Embeddings-based methods (Emb). In this widely used (Liu et al., 2021; Periti and Montanelli, 2024, inter alia) family of approaches, word usage is first captured by a contextualized embedding vector. From there, a word's representations from all sentences within one time period are aggregated into a prototype vector **(PRT)**; change is computed as the distance between two prototype representations (Martinc et al., 2020). Alternatively, change can be computed as the average pairwise distance between embedding vectors from different time periods **(APD)** (Kutuzov and Giulianelli, 2020).

Sense/Cluster-based methods (Clstr). In another common type of semantic-change detection method (Martinc et al., 2020; Montariol et al., 2021, inter alia), the contextualized embeddings of for different occurrences of the same word are grouped into clusters that are further refined to represent different senses of the word in the period. We use Affinity Propagation **(AP)** and K-means clustering **(K5)**. Period representations are obtained as the distribution of word uses across these sense clusters, and semantic change is quantified as the Wasserstein distance (Solomon, 2018) between two period distributions.

Substitutes-based methods (Subst). Introduced by Card (2023), this method represents word usage instances as the top-5 lexical-substitute candidates obtained from a masked-token prediction task. Representations are then aggregated as the distribution over all possible substitutes, and semantic change is quantified as the Jensen–Shannon divergence score between two period distributions. Similarly, Periti et al. (2024) use substitution to examine how language models contextualize words in different usage contexts.

5.4 Secondary Techniques

Further, we also evaluate techniques for improving semantic change detection by incorporating additional control measures. Because, at its core, introduction of control measures helps distinguish genuine changes from erroneous ones due to noise in the data, we expect these techniques to aid SCDisc.

Permutation Tests. (PT) Some previous research turned to statistical tests to improve the performance of semantic-change quantification methods (Kulkarni et al., 2015; Liu et al., 2021). These tests determine whether shifts captured by semantic-change detection methods are statistically significant or not, essentially identifying false positives among the detected changes. Hence, applying them to filter through $\hat{f}$-sorted list is expected to aid the SCDisc setting, too.

For our experiments, we select the statistical tests introduced by Liu et al. (2021). Here, we obtain the distribution of possible change scores by shuffling use cases of the word w between two time periods and then determining whether the detected semantic change δ_w is statistically significant. Like the authors, we also complement permutation tests with a False-Discovery-Rate **(PT + FDR)** control technique (Benjamini and Hochberg, 1995).

We apply permutation tests to results from our embeddings- and substitutes-based methods; we

could not afford the computational cost of running permutation tests in combination with Montariol et al. (2021) clustering.

Scaled Change Metrics. Previous research has observed that change scores may be biased towards high vs. low frequency words (Card, 2023) or words that appear in more variable contexts (Noble et al., 2021). Therefore, these researchers define adjusted metrics to control for such biases and improve semantic-change detection.

In our experiments, we start with frequency-based scaling **(FS)** (Card, 2023). For each word, the scaled metric represents how much it has changed compared to other words in the same frequency range. For a word $w \in V$ consider a set of terms "comparable" to w in terms of frequency[8]: $S_w = \{x \in V \mid \text{freq}(w)/F \leq \text{freq}(x) \leq \text{freq}(w) * F\}$ for some predetermined constant F. Then, the scaled change score is computed as $\mu(\mathbb{I}_x[\delta_w \geq \delta_x])$, where μ is mean, δ_w and δ_x are raw semantic-change scores for w and $x \in S_w$, respectively. We further extend this scaling technique to also control-match for part-of-speech information between w and $x \in S_w$ **(FS + PM)**.

6 Results

6.1 Ranking Evaluation Results

Base-methods. When the methods are run without secondary techniques, the average rank induced on the items in T^* is disappointingly low. Consult the rows in Table 2 without a ↳ in the label:[9] while the perfect value for either corpus would be 15/2 (since $|T^*| = 15$), we see T^*'s average rank being *at best* 351 (LiverpoolFC) and 3405 (SemEval-EN): hundreds of terms from V receive better $\hat{f}$ scores than some of the terms in T^*. Besides the aggregated T^* described above, we also show with the empty-icon lines in Figure 2 "when" in a base method's ranking a new T^* word is discovered, i.e., how $\phi(r)$ varies with r: they are all unfortunately distant from the ideal $y = x$ line.

Scaled Change Metrics. Across practically all model & base-method combinations and both corpora, scaling $\hat{f}$ improves the induced average rank of T^*. This is depicted in Figure 3 by arrows leading from an empty icon (base method) to a shaded icon (+FS version) and then to a darker-shaded icon (+FS+PM version). The first "hop" is generally towards the optimal result (yellow star, top-right), more so along the LiverpoolFC axis. One can consult Figure 5 for a zoomed-in visualization (where it is apparent that FS outperforms FS+PM), but Figure 3 clearly shows that the improvements delivered by scaling still do not produce reasonable ranking results for *all* of T^*.

It is true, as shown by comparing the filled- vs. open-icon lines in our plots of $\phi(r)$ (Figure 2), that *some* T^* words are better-placed by FS — indeed, in Figure 2b, XLM-R Subst (JSD) + FS initially traces the $y = x$ ideal line. But, recall that T consists of the *very top* 15 highest-$\ell(\cdot)$ words with respect to the original benchmark annotations, and thus *all* should receive relatively good ranks. Unfortunately, we see that roughly a third of T^* gets placed at ranks ≈ 100 or worse, sometimes even in the ten-thousands.

Permutation tests. We see mixed results with both permutation tests and the added false-discovery-rate correction (↳ PT and ↳PT +FDR words in Table 2). When things go well, they do move T^* upward in the rankings. However, sometimes they may also mark some words in T^* as having no statistically significant change, bounding $\phi(r) < 1$ for all r. For instance, in LiverpoolFC with BERT Emb (PRT) + PT four words of T^* didn't pass the statistical significance test, dropping the average rank to 3396 from the original 547 with BERT Emb (PRT). Generally, we see permutation tests do worse in LiverpoolFC than in SemEval-EN.

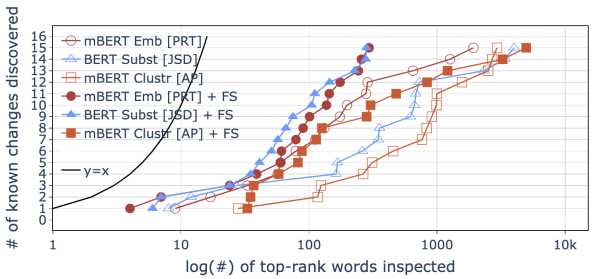
6.2 Discovery Validation via Human Annotations

As noted in the Introduction, we do need to check whether the words in $V \setminus T$ — which initially lack labels — that receive better $\hat{f}$-induced ranks than words in T^* might actually be true semantic changes. We therefore apply additional human annotations and $\psi(k)$ to gain further understanding of SCDisc performance than what is possible with $\phi(r)$.

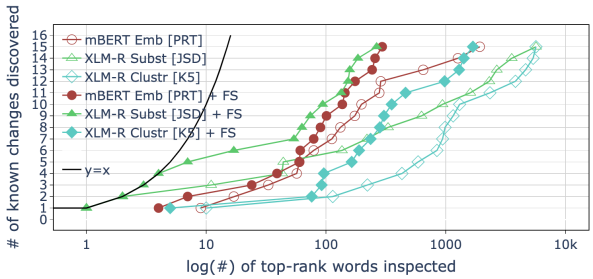
For SemEval-EN, despite generally poor ranking performance, human annotations show that the majority of $W'[:15]$ across all annotated methods are in fact semantic changes (see Table 1). This corroborates observations by Kurtyigit et al. (2021) for semantic-change discovery in a comparable German SemEval dataset. BERT Emb (PRT) and

[8]The frequency bounds were selected based on the best semantic-change detection performance on T.

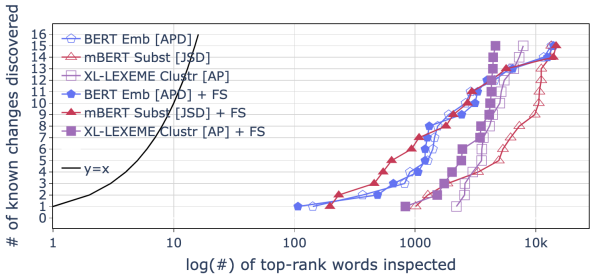
[9]Also visualized after $|V|$-normalization as the empty icons in Figure 3.



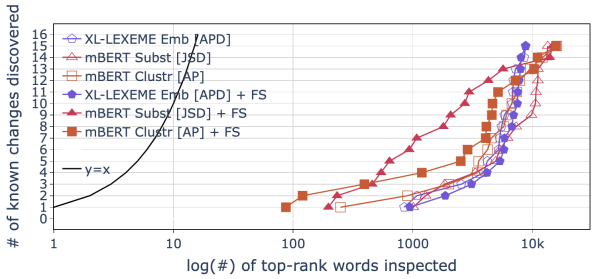
(a) LiverpoolFC: best avg rank



(b) LiverpoolFC: best first rank



(c) SemEval-EN: best avg rank



(d) SemEval-EN: best first rank

Figure 2: Discovery performance in LiverpoolFC and SemEval-EN: $\phi_{\hat{f}}(r) * |T^*|$ with respect to $log(r)$ on the x-axis across different methods, in their base implementation and with frequency-based scaling. For each of the three representation types, we select model & base-method combinations according to either the average rank of T^* (a & c) or the rank of the first discovered element of T^* (b & d) achieved with the base-method. The $y = x$ line indicates the "ideal" performance where the top changes are exactly T^*.

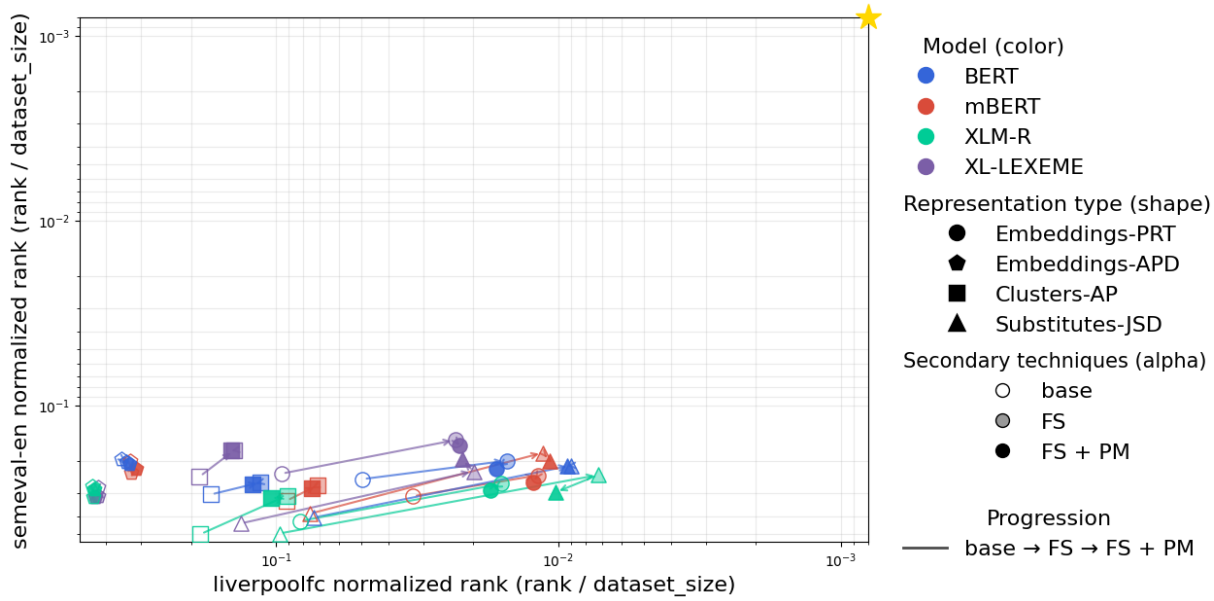


Figure 3: Average normalized rank for T^* in the SemEval-EN/LiverpoolFC performance plane across 16 model-representation combinations. The yellow star indicates the optimal location (top-right corner). Arrows indicate how performance changes as we apply secondary techniques.

BERT Clstr (AP) (as base-methods) yield best results. Overall, secondary techniques don't improve $\psi(k)$ in SemEval-EN.

By contrast, in LiverpoolFC for most methods, less than a half of $W'[15]$ are annotated as changed in meaning. Although as a base-method Emb (PRT) produced best $\phi_{\hat{f}}(r)$ results, through the lens of human annotations, it performs worst among other representation types: only three words with semantic change are found in $W'[15]$. In contrast

	SemE. (↑)	Liverp. (↑)
Embeddings-based	93.3	20.0
↳+ FS	93.3	40.0
↳+ FS + PM	86.7	26.7
↳+ PT	73.3	53.3
↳+ PT + FDR	86.7	40.0
Clusters-based	93.3	53.3
↳+ FS	93.3	46.7
↳+ FS + PM	80.0	53.3
Substitutes-based	80.0	40.0
↳+ FS	80.0	20.0
↳+ FS + PM	80.0	26.7

Table 1: $\psi(k)$ (%) with $k = 15$ across different semantic change quantification methods: percentage of words among $W'[:k]$ that were verified to have changed in meaning by human annotations. We report for three types of methods [**BERT Emb (PRT)**, **BERT Subst**, **BERT Clstr (AP)**] and also for metrics scaled with respect to words with similar frequencies **(FS)**; metrics scaled with respect to frequencies in addition to matched part-of-speech **(FS + PM)**

to SemEval-EN, frequency scaling (FS) resulted in better values of $\psi(k)$. For other methods, however, FS and FS+PM do not lead to much improvement in discovery performance. Surprisingly, permutation tests PT led to even more semantic changes ranked among $W'[15]$. Finally, human annotations show that 8 out of 15 words highly ranked by Clstr (AP) have in fact change in meaning, which is the best $\psi(k)$ achieved by a base-method in LiverpoolFC.

7 Discussion and Conclusion

7.1 LiverpoolFC vs SemEval-EN

SCDisc performance varies greatly between SemEval-EN and LiverpoolFC, and we hypothesize this discrepancy to be rooted in the properties of their corpora.

Despite their poor ranking performance in SemEval-EN, semantic-change detection methods do great when evaluated using $\psi(k)$. This means that there *are* quite a few words in V that change in meaning by just as much (if not more) than words in T^*. This is partially possible due to time periods in SemEval-EN being far apart, thus allowing greater room for change. At the same time, we only considered $\psi(k)$ at $k = 15$, and one may still be skeptical of whether thousands of words ranked above T^* in SemEval-EN are indeed genuine changes.

On the other hand, semantic-change detection methods on LiverpoolFC perform slightly better at ranking T^*, but struggle with false positives at the top of $\hat{f}$-sorted lists. False positives may be a result of variations in usage contexts of words and sensitivity of methods to them. It is possible that for some words in V, a time slice sub-corpus C_i has a limited number of examples where these words appear in *varied contexts* but carry the *same meaning*. Thus, observed variation in context (even without underlying meaning change) could falsely be equated to semantic change by methods. This is in some way similar to what happens with "data bursts" in Kutuzov et al. (2022). Perhaps therefore, with more sensitive Emb (PRT) representations, frequency scaling (FS) stabilizes semantic-change scores by comparing words similar in frequency and in the process improves both $\phi(r)$ and $\psi(k)$ (Table 2).

At the same time, variations in usage context are sometimes indicative of changing trends in the world and within the community, adoption of a meme, or emergence of expressions reflecting some shared in-group understanding — and these *are* instances of short-term semantic change (Del Tredici et al., 2019). So, slightly better $\phi(r)$ (compared to SemEval-EN) may be attributed to such context variability patterns in T^* itself. Although short-term semantic change is different from its traditional long-term counterpart, methods should excel both at *quantifying* and at *discovering* new instances of such change.

7.2 Recall vs Precision

We investigate two approaches for evaluating semantic-change detection methods on SCDisc: ranking-based $\phi(r)$ and annotations-based $\psi(k)$. One may view the former as Recall: $\phi(r)$ measures how well semantic-change detection methods identify (via ranking up to r) known highest changes. Similarly, the latter corresponds to Precision: $\psi(k)$ measures how many of identified changes (within $W'[:k]$) are true positives. Good semantic-change detection methods should do well at both Precision and Recall of the SCDisc task. Hence, it is important that researchers introducing new methods also evaluate them for SCDisc by [i] considering their rankings of T^* and [ii] annotating according to a top-k approach on data with a natural distribution of changed vs. non-changed words.

8 Limitations

We evaluate SCDisc on two datasets: SemEval-EN is a carefully curated benchmark of historical semantic changes, while LiverpoolFC presents a unique dataset of short-term semantic changes in the wild. We draw observations about the SCDisc performance of semantic-change detection methods only by contrasting results observed in these two contexts. Hence, for a more comprehensive view of semantic-change discovery future research should evaluate SCDisc in more datasets. These should be both semantic change detection benchmarks and other (perhaps previously unannotated) corpora. In addition to long-term semantic change, it is also important to consider more corpora where one can observe instances of short-term semantic change.

Due to limited ability of collecting annotations in other languages, we only evaluate SCDisc on English corpora. Hence, future research should examine capabilities of semantic-change detection methods at semantic-change discovery in other languages.

Although we design our annotation task to be similar to Del Tredici et al. (2019), due to factors such as time passed since the benchmark publication, a selection of a different set of annotators, and variations in phrasing of the task and instructions, it is possible that our annotations are different from the original ones. We try to mitigate this effect by also annotating the 5 highest-change words and 5 non-changes in T, and comparing our annotations for them. Generally, our non-change annotations match; we miss one of the highest-change words. However, beyond this limited control set, we are unable to make judgments of closeness of our annotations to those by Del Tredici et al. (2019).

On the other hand, for SemEval-EN our annotation procedure is substantially different from the one used by Schlechtweg et al. (2020). Even though we verify our annotations using 5 highest-change words and 5 non-changes from T (like we did with LiverpoolFC), they are still quite distinct from the other annotations of the benchmark.

Furthermore, for consistency with LiverpoolFC, in data processing, obtaining contextualized model representations, and quantifying change across all methods, we use "token" format of available SemEval-EN data. Next, to run ranking evaluations, we convert T of SemEval-EN from lemmas into words (types) by running a sentence-level matching algorithm and finding all words that map to each lemma in the original T. We observe that some words mapped to these lemmas have few appearances in C, and hence may show little variance in their contextualized representations across time periods. This, in addition to the fact that the original $\ell(t)$ were provided for lemmas, could be a contributor to poor performance of methods when ranking T^*. Therefore, future research should explore semantic-change discovery at the lemma-level in SemEval-EN or other datasets.

9 Acknowledgments

We thank the anonymous reviewers for particularly insightful comments and important suggestions; we are grateful for your attention and care with our submission! We thank Jack Hessel, Jacob Matthews, and Qi Han for helpful conversations and inspiring discussions. We also express our gratitude to the team of annotators whose meticulous work and efforts were instrumental to our project. Thank you Annabelle Ford, Audrey Kim, Carlos Alvarez, Donna Gan, Elias Decker, Emely Alvarez, Khondamir Umarov, Lynandrea Mejia, Malika Inomzoda, and Pado Roberts. This work was supported in part by a gift from Google. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Google.

References

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

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, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-Term Meaning Shift: A Distributional Exploration. 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 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Copenhagen, Denmark. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

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

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.

Sinan Kurtyigit, Maike Park, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Lexical semantic change discovery. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6985–6998, Online. Association for Computational Linguistics.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

Andrey Kutuzov and Lidia Pivovarova. 2021. Three-part diachronic semantic change dataset for Russian. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 7–13, Online. Association for Computational Linguistics.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022. Contextualized embeddings for semantic change detection: Lessons learned. *Northern European Journal of Language Technology*, 8.

Yang Liu, Alan Medlar, and Dorota Glowacka. 2021. Statistically Significant Detection of Semantic Shifts using Contextual Word Embeddings. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 104–113, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

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

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

Bill Noble, Asad Sayeed, Raquel Fernández, and Staffan Larsson. 2021. Semantic shift in social networks. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 26–37, Online. Association for Computational Linguistics.

Francesco Periti, Pierluigi Cassotti, Haim Dubossarsky, and Nina Tahmasebi. 2024. Analyzing semantic

change through lexical replacements. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4495–4510, Bangkok, Thailand. Association for Computational Linguistics.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Computing Surveys*, 56(11).

Francesco Periti and Nina Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico. 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, 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.

Justin Solomon. 2018. Optimal transport on discrete domains. *AMS Short Course on Discrete Differential Geometry*.

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.

A Human Annotations

Prior to the task the annotators were presented with the following instructions: `Semantic change occurs when the meaning of words change over time. In this annotation task we are going to provide a set of words and ask to annotate for changes in their meaning. For each word, you will be presented with two groups of sentences, in which the word occurs, and asked to indicate whether or not there is change in meaning of the highlighted word between sentences in Group 1 and Group 2. We are interested in human perceptions of language change, so we ask not to use generative AI assistants when making your judgments.`

Annotators were supplied with a .xlsx document, where they would provide annotations. Each sheet in the document would correspond to one word. Each word was accompanied with 7 questions, which we list below ("****" marks the word that is being annotated).

- Q1 Is there a primary meaning (i.e., majority sense) for "****" in Group 1? (YES/NO)
- Q2 Is there a primary meaning (i.e., majority sense) for "****" in Group 2? (YES/NO)
- Q3 Is there a difference in the majority sense of "****" in Group 1 and the majroity sense of "****" in Group 2? (YES/NO). Mark N/A if answered NO to Q1 or Q2.
- Q4 Was it difficult to answer Q3? (YES/NO). Mark N/A if answered NO to Q1 or Q2.
- Q5 Please provide sentence #s if change in meaning is present. If no change was detected, mark N/A.
- Q6 Is there a sense/meaning of "****" in Group 1 that is not present in Group 2? If YES, indicate sentence #. If NO, mark as N/A.
- Q7 Is there a sense/meaning of "****" in Group 2 that is not present in Group 1? If YES, indicate sentence #. If NO, mark as N/A.

B Additional Ranking Results

Sent #	Group 1	Group 2
S1	whether we **accomplish** that this season is the question .	i fully believe in what he is trying to **accomplish** here , even if i don ' t fully understand it .
S2	i don ' t really understand what dalglish was trying to **accomplish** yesterday with the squad he put out but if he picks the same side next week in that formation arsenal ' s midfield , even without cesc and nasri , will probably punish us .	although meme magic may be able to **accomplish** that .
S3	i do agree with everyone that he should stay for one more season and if we don ' t **accomplish** a top four finish , then he has the right to speak with other clubs .	yeah , i ' m sure this will **accomplish** that .
S4	to **accomplish** our long - term goals .	if we ever want to **accomplish** anything and not have that one odd season where we come close to 1st than we need to think bigger than one players demands , i don ' t care if he wants to go , he has a contract until 2022 .
S5	whereas the same if done by avb , would have been considered as a failure to **accomplish** what he was hired for .	tugged in multiple directions yet goes no where and generally doesn ' t **accomplish** much
Q1:	Is there a primary meaning (i.e., majority sense) for "**accomplish**" in Group 1? (YES/NO)	
Q2:	Is there a primary meaning (i.e., majority sense) for "**accomplish**" in Group 2? (YES/NO)	
Q3:	Is there a difference in the majority sense of "**accomplish**" in Group 1 and the majroity sense of "**accomplish**" in Group 2? (YES/NO). Mark N/A if answered YES to Q1 or Q2.	
Q4:	Was it difficult to answer Q3? (YES/NO). Mark N/A if answered YES to Q1 or Q2.	
Q5:	Please provide example sentence #s if change in meaning is present. If no change was detected, mark N/A.	
Q6:	Is there a sense/meaning of "**accomplish**" in Group 1 that is not present in Group 2? If YES, indicate sentence #. If NO, mark as N/A.	
Q7:	Is there a sense/meaning of "**accomplish**" in Group 1 that is not present in Group 2? If YES, indicate sentence #. If NO, mark as N/A.	

Figure 4: The screenshot of the annotation spreadsheet for a sample word "accomplish". Red circle marks the area for providing annotations.

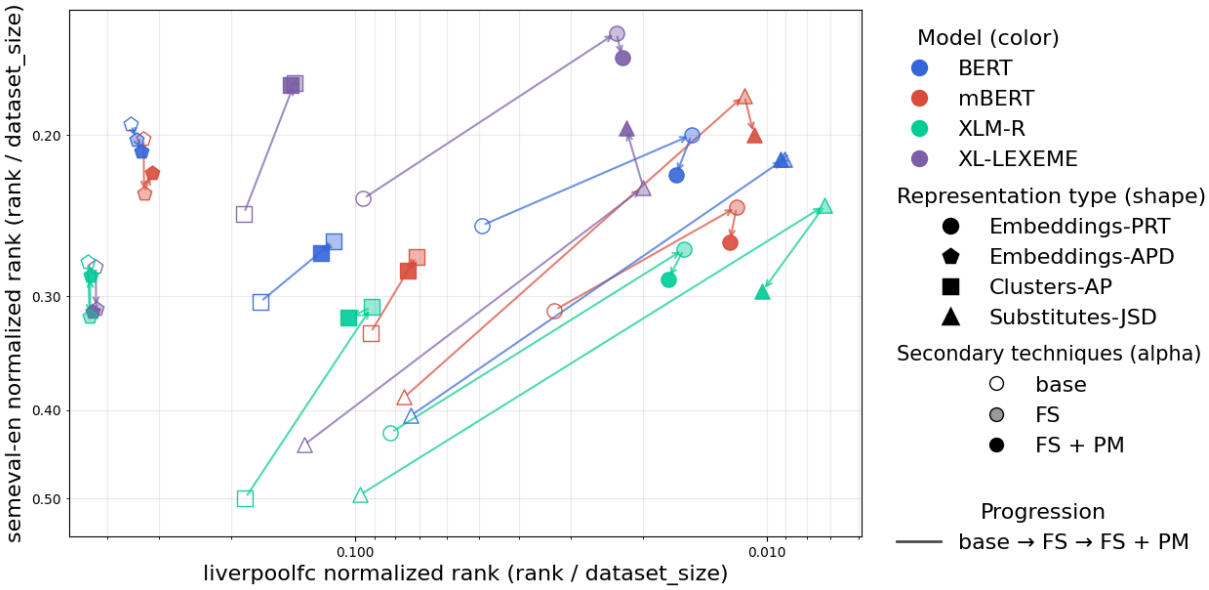


Figure 5: Zoomed-in version of Figure 3 (average normalized rank for T^* in the SemEval-EN/LiverpoolFC performance plane): the optimal location is now off the figure beyond the top-right corner.

	SemEval-EN avg rank (↓)				LiverpoolFC avg rank (↓)			
	BERT	mBERT	XLM-R	XL-LEXEME	BERT	mBERT	XLM-R	XL-LEXEME
Emb (PRT)	4412	5698	7824	**4308**	547	**351**	914	1125
↳+ FS	3459	4327	4833	**2813**	144	**113**	145	254
↳+ FS + PM	3981	4835	5429	**3089**	183	**137**	196	265
↳+ PT	617	**613**	1765	1695	3396	**1772**	3370	4960
↳+ PT + FDR	**497**	1697	1476	2692	**5701**	6489	7275	8878
Emb (APD)	**3405**	3667	4926	5302	4391	**4008**	5240	5315
↳+ FS	**3539**	4215	5691	5888	4252	**4004**	5184	5310
↳+ FS + PM	**3808**	4080	5389	5995	4200	**3896**	5406	5478
↳+ PT	3988	4086	**3928**	8830	**5703**	7368	8886	10484
↳+ PT + FDR	3978	4077	**3920**	10054	**5703**	8143	9678	10484
Clustr (AP)	5500	6026	9302	**4476**	2046	**1033**	2121	2276
↳+ FS	4707	4937	5564	**3222**	1332	**793**	949	1704
↳+ FS + PM	4156	3894	5965	**3833**	**112**	127	129	261
Clustr (K5)	**4913**	7182	10150	5566	**1551**	1603	1896	2469
↳+ FS	**3312**	4569	5812	4209	**427**	495	508	1706
↳+ FS + PM	4998	5254	5980	**3376**	1477	**886**	1221	1798
Subst (JSD)	7798	**7409**	9588	8259	**916**	935	1211	1602
↳+ FS	4150	**3487**	4774	4302	100	120	**80**	215
↳+ FS + PM	**3718**	5078	6479	4356	**500**	505	745	1894
↳+ PT	**661**	1840	2953	1692	5672	4159	**3344**	5771
↳+ PT + FDR	1610	1579	2683	**1471**	5672	6482	**4861**	5679

Table 2: Average rank of T^* in W of SemEval-EN and LiverpoolFC, datasets, presented for various combinations of models, methods, and secondary techniques. The best value per dataset per row is in **bold**; a box indicates the best result per "rectangle" = fixed method, fixed corpus, various models. .