Literary Intertextual Semantic Change Detection: Application and Motivation for Evaluating Models on Small Corpora

Jackson Ehrenworth
Williams College
jne1@williams.edu

Katherine A. Keith
Williams College
kak5@williams.edu

Abstract

Lexical semantic change detection is the study of how words change meaning between corpora. While Schlechtweg et al. (2020) standardized datasets and evaluation metrics for this shared task, for those interested in applying semantic change detection models to small corpora—e.g., in the digital humanities—there is a need for evaluation involving much smaller datasets. We present a method and open-source code pipeline for downsampling the SemEval-2020 Task 1 corpora while preserving gold standard measures of semantic change. We then evaluate several high-performing unsupervised models on these downsampled corpora, and find that the models experience both dramatically decreased performance (average 67% decrease) and high variance. Finally, we propose a novel application to the digital humanities: *literary intertextual semantic change detection*, the production of a ranked list of words by degree of semantic change between two books. We then provide a case study of this application to Fanon's *The Wretched of the Earth* and Hartman's *Scenes of Subjection* and find that semantic change detection models—even with their current limited performance on small corpora—may still produce fruitful avenues of exploration for literary scholars.

1 Introduction

Semantic meaning is fluid. The word *plane*, for instance, underwent a dramatic semiotic shift around the early 1900s from the sense of "flat geometric surface" to the sense of "aeroplane" (oed). The last ten years have seen the rise of computational linguistic approaches that attempt to provide unsupervised detection of lexical semantic change (Kutuzov et al., 2018; Tahmasebi et al., 2021). Applications include discovering laws of semantic change (Xu and Kemp, 2015; Hamilton et al., 2016b; Dubossarsky et al., 2017; Boleda, 2020), investigating the evolution of harmful stereotypes (Garg et al., 2018), or determining how societal relationships to certain concepts experience diachronic drift (Kozlowski et al., 2019), among others.

The majority of these fields involve studying *large* conglomerate corpora as proxies for societal beliefs. In the burgeoning literary digital humanities (Gold, 2012; Kirschenbaum, 2016; Eve, 2022), among other fields, however, one is often invested in studying *small* corpora, where each corpus is on the order of 150k tokens (about the size of a single authored English fiction novel). Schlechtweg et al. (2020) standardized evaluation metrics and datasets for unsupervised semantic change detection, but Schlechtweg et al.'s smallest corpus contains over 1.7 million tokens, and their largest over 110 million. In this work, we investigate the degree of performance degradation of semantic change detection models when evaluated on small corpora. We expect this setting to be challenging for the evaluated models due to the limited number of examples of each target word in context available to them.

To further motivate the importance of evaluating semantic change detection models on small corpora, we focus on applying these models to aid literary studies. In the context of literary criticism, investigating subtle differences in language between two books often provides the building blocks for broader comparative literary insight. In this work, then, we propose a novel application—*literary intertextual semantic change detection*, the production of a ranked list of words by degree of semantic change between two books—as an exploratory tool to suggest words that may be of literary interest and suitable for extended investigation (e.g., through comparative close-reading by humans). In this setting, corpora sizes are limited by the length of the books under consideration.

Finally, as a case study for how, and, importantly to us, whether, current semantic change detection models can be employed to produce fruitful avenues of inquiry for literary scholars, we apply the best performing English model evaluated in Sec-

Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change, pages 1–14

tion 5.1 to two books—*The Wretched of the Earth* (Fanon, 1961/2021) and *Scenes of Subjection* (Hartman, 1997/2022)—which we suspected may have interesting intertextual semantic changes due to prior domain knowledge. We find that there is reason to be optimistic that semantic change detection can be used in an exploratory manner to aid literary critics.

To summarize, our primary contributions are the following:

- We create an evaluation framework that enables the downsampling of the SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection datasets presented by Schlechtweg et al. (2020) while preserving ground truth data.[1]
- We evaluate a few of the best-performing semantic change detection models on downsampled corpora and find both dramatic decreases in performance (average 67% decrease) and high variance, opening the door to future work building models specifically for this low-resource setting.
- We propose a novel application of semantic change detection to the digital humanities—*literary intertextual semantic change detection*—and, through a case study of two books (*The Wretched of the Earth* (Fanon, 1961/2021) and *Scenes of Subjection* (Hartman, 1997/2022)) demonstrate the usefulness of these types of models for literary criticism.

2 Related Work

2.1 Methods for Semantic Change Detection

Methods for semantic change detection can be loosely categorized into four groups, the majority of which use cosine similarity between word embeddings created from two corpora as a proxy for semantic change. First, there are count-based methods that rely on explicit co-occurence matrices or their derivatives (Sagi et al., 2009; Cook and Stevenson, 2010; Gulordava and Baroni, 2011). There has been a general shift away from these initial methods towards the use of prediction-based models—such as those based on Continuous Skip gram with negative sampling (Mikolov et al., 2013, SGNS)—for the creation of word embeddings, with various strategies for aligning embeddings across time steps (Kim et al., 2014; Kulkarni et al., 2015; Dubossarsky et al., 2019). Recently, the use of contextualized word embeddings, derived predominantly from the BERT (Devlin et al., 2019) or ELMo (Peters et al., 2018) architectures, have seen a surge in popularity in the field. Generally, contextualized word embeddings are created by fine-tuning pre-trained language models on the corpora under consideration and then extracting and clustering or averaging hidden layer weights (Giulianelli et al., 2020; Martinc et al., 2020; Montariol et al., 2021; Rosin et al., 2022; Rosin and Radinsky, 2022). Separately, there are also probabilistic or dynamic methods that use both context-free (Bamler and Mandt, 2017; Rosenfeld and Erk, 2018) and context-based (Hofmann et al., 2021) mechanisms.

2.2 Applications of Word Embeddings to Small Corpora Tasks

While we are not aware of any research quantitatively evaluating semantic change detection models on small corpora, word embeddings that perform well on tasks involving small corpora have applications to, and have been studied in, a variety of fields.[2] Word embeddings have been used in psychology to detect formal thought disorder in transcribed or written statements (Voleti et al., 2020; Sarzynska-Wawer et al., 2021) and studied for their ability to capture word associations in dream reports (Altszyler et al., 2017; Elce et al., 2021). In the field of philosophy, meanwhile, domain-expertise has been used to investigate whether word embeddings can cluster related concepts in large single authored corpora with domain-specific content (Betti et al., 2020; Oortwijn et al., 2021). And political scientists have developed methods to support significance testing for use in contexts where corpora are large but target words are domain-specific and generally rare (Rodriguez et al., 2023).

Despite the broad interest in investigating the ability of word embeddings to capture semantic meaning even when data is scarce—a literature that this paper compliments—we are not aware of any attempts to evaluate the approaches surveyed in Section 2.1 on semantic change detection tasks for small corpora, nor are we aware of any annotated

[1] All experiments and code are available at https://github.com/jnehrenworth/small-corpora-scd.

[2] While Montariol and Allauzen (2019) have studied semantic change detection models in the context of scarce data, their research occurred before Schlechtweg et al. (2020) and, because of this, was limited to empirical evaluation on corpora without gold-standard data.

test sets for semantic change detection covering corpora small enough to simulate single books.

2.3 Applications of Semantic Change Detection to the Digital Humanities

Semantic change detection applied to the digital humanities is still nascent. Nevertheless, this intersection has previously been hinted at as a direction for future work by authors working in the field of semantic change detection (Tahmasebi and Risse, 2017; Kutuzov et al., 2018; Tahmasebi et al., 2021), and there is other prior work at this intersection. Semantic change detection has been used to track semantic innovation in abolitionist newspapers (Soni et al., 2021), investigate a debate about compositional shifts in a single authored series of Danish historical works (Nielbo et al., 2019), study evolving representations and stereotypes of Jewish people in 19th century France (Sullam et al., 2022), track the transformation of tropes in a large curated corpus of German poetry (Haider and Eger, 2019), and attempt to model character relations in the *Harry Potter* series (Volpetti et al., 2020; K et al., 2020).

While the exploratory use of semantic change detection in the digital humanities is not novel, we are not aware of any papers that suggest using semantic change detection directly to produce a ranked list of words by intertextual semantic change as an avenue for comparative literary analysis.[3]

3 Datasets

3.1 Overview

The standard datasets and shared tasks for semantic change detection were presented by Schlechtweg et al. (2020) in "SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection." Schlechtweg et al. (2020) released corpora in four languages—English, German, Latin, and Swedish—each of which are bifurcated diachronically at some time period. The released corpora are genre-balanced year to year. Abbreviated summary statistics of these corpora are given in Table 1.

For each pair of corpora in a given language, call them C_1 and C_2, Schlechtweg et al. (2020) present two subtasks: 1. binary classification, and 2. ranking the degree of semantic change. Our work is

[3]Our proposed application can be considered a near special-case of Lexical Semantic Change Discovery (Kurtyigit et al., 2021), except that our use-case is focused on graded change rather than that of binary classification (see Section 3.1 for more details).

	C_1 Tokens	C_2 Tokens	Target Words
English	6.5M	6.7M	37
German	70.2M	72.3M	48
Swedish	71.0M	110.0M	31

Table 1: Summary statistics for SemEval-2020 Task 1 corpora, abbreviated from Schlechtweg et al. (2020). C_1 and C_2 are time-specific corpora. *Target Words* indicate the number of evaluation words to be ranked by degree of semantic change between C_1 and C_2.

	C_1			C_2		
	Gold	Random	Total	Gold	Random	Total
English	138k	12k	150k	94k	56k	150k
German	156k	0k	156k	125k	25k	150k
Swedish	107k	43k	150k	95k	55k	150k

Table 2: Summary statistics for downsampled corpora, where: *Gold* is the number of tokens selected from lines used in the manual annotation process, as found in the usage graphs of Schlechtweg et al. (2021), *Random* is the number of tokens from lines randomly sampled until 150k total tokens were included, and *Total* is the total number tokens included.

exclusively focused on Subtask 2, where the the goal is to determine the amount of semantic shift that a list of target words have undergone between C_1 and C_2 by proxy of ranking them according to their degree of semantic shift (e.g., "gay" has changed more than "cell" which has changed more than "peer"). For one, it seems intuitively likely (and the results presented by Schlechtweg et al. (2020) tend to bear this out) that high performance on Subtask 2 is indicative of high performance on Subtask 1. It also seems to us as if Subtask 2 captures more about the subtle movement of language that literary critics are generally interested in. For instance, a polysemous word may not experience a binary sense change between C_1 and C_2 while still shifting from primarily one sense type to another. The production of a ranked list of words carries another, perhaps ancillary, benefit: it provides a literary critic the ability to easily prioritize which words to investigate more thoroughly. We believe this ability to be especially relevant because of how onerous we found it in our case study (Section 6) to determine for a given word: a) what the semantic change was, and b) whether the semantic change had literary relevance.

For each language, Schlechtweg et al. (2020)

released gold standard data for a subset of target words balanced for part of speech and frequency: via a manual annotation process, each target word was assigned a label between 0 and 1 denoting degree of semantic change (0 means no change has taken place, 1 is the maximum amount of change).

3.2 Downsampling Method

This paper focuses on downsampling the datasets presented by Schlechtweg et al. (2020) while preserving the gold standard data obtained via manual annotation. The annotation process, described in detail by Schlechtweg et al. (2021), involved selectively annotating pairs of word uses to create a sparsely connected usage graph. As randomly sampling a certain number of sentences from the SemEval-2020 Task 1 corpora until a target token amount is met would destroy this usage graph, we preserved it via the following steps:

1. After pre-processing and cleaning the text (see Appendix A), we used exact matching, to cross-reference the context text of each raw annotated use—presented in Schlechtweg et al. (2021)—used to create the SemEval-2020 Task 1 gold standard data with its counterpart in the SemEval-2020 Task 1 corpora (Schlechtweg et al., 2020).[4]

2. We programmatically selected all lines from the SemEval-2020 Task 1 corpora that were part of the manual annotation process.

3. We then took a random sample of additional lines until a desired token threshold was reached.

For the experiments presented in this paper, a token threshold of 150k was used. The German C_1 corpus had 156k tokens already present from the annotated sentences, so no additional random sampling occurred. For summary information about the downsampled corpora see Table 2.

4 Evaluating Existing Methods on Small Corpora

In this paper we evaluate three models that present a range of different architectures, from static (non-contextual) embeddings to contextual embeddings, and are, to our knowledge, among the highest performing open-source models for unsupervised semantic change detection:

1. Pražák et al. (2020), the winning submission on SemEval-2020 Task 1, Subtask 1. The authors train static (non-contextual) embeddings using SGNS, align them using orthogonal Procrustes, and then use cosine distance to compare aligned embeddings.

2. Pömsl and Lyapin (2020), the winning submission on SemEval-2020 Task 1, Subtask 2. The authors train static (non-contextual) embeddings using SGNS, align them using orthogonal Procrustes, and then take Euclidean distance as their metric when comparing aligned embeddings.[5]

3. Rosin and Radinsky (2022), the highest performing open-source contextualized semantic shift detection model on Subtask 2 we are aware of (Montanelli and Periti, 2023). The authors propose a temporal self-attention mechanism as a modification to the standard transformers architecture. They use a pre-trained BERT model, fine-tune it on diachronic corpora using their proposed temporal attention mechanism, and then create time-specific representations of target words by extracting and averaging hidden-layer weights. These representations are then averaged at the token level and compared using cosine similarity.[6]

We have used the models essentially as-is from their respective GitHub repositories. Hyperparameters for all models were chosen based on those reported in each paper. Note that both Pražák et al. (2020) and Pömsl and Lyapin (2020) learn static (non-contextual) embeddings from scratch on the target corpora, while the contextualized model of Rosin and Radinsky (2022) is already pre-trained and only fine-tuned on the target corpora.

[4] We used the lemmatized versions of both the SemEval-2020 Task 1 corpora and the annotated uses for this matching procedure. Due to larger inconsistencies in formatting between the Latin annotated uses and the SemEval-2020 Task 1 corpora, we were unable to successfully devise a way to cross-reference Latin annotated uses (see Appendix A). For that reason, the Latin corpora was excluded from this study.

[5] Note that although Pömsl and Lyapin describe ensemble and models with contextualized embeddings in their paper, their winning submission used static (non-contextual) embeddings and is what we have chosen to evaluate.

[6] For the purposes of this study, we use the best tested version of BERT (Devlin et al., 2019) for each language from HuggingFace's repository, as reported by the authors (`bert-tiny` for English and `bert-base-german-cased` for German).

Model	SemEval-Small Avg.	EN	DE	SV	SemEval Avg.	EN	DE	SV	Δ
Pražák et al. (2020)	0.269	0.106	0.361	0.340	0.481	0.367	0.697	0.604	−44%
Pömsl and Lyapin (2020)	0.049	0.060	0.022	0.066	0.527	0.422	0.725	0.547	−90%
Rosin and Radinsky (2022)	0.226	0.320	0.132		0.695	0.627	0.763		−67%

Table 3: Summary view of mean performance across 500 downsampled corpora (SemEval-Small), measured using Spearman's ρ, along with best performance as reported by Schlechtweg et al. (2020) or by Montanelli and Periti (2023) (SemEval). Δ refers to average percent decrease in performance between the SemEval corpora and the downsampled corpora, while EN, DE, and SV denote performance on the English, German, and Swedish corpora, respectively.

We do not intend for this to be an exhaustive evaluation of all possible methods in the field. Instead, we hope to open the door for future research to evaluate other methods for semantic change detection — perhaps Nonce2Vec (Herbelot and Baroni, 2017), LSA+SVD (Deerwester et al., 1990), or PPMI+SVD (Levy et al., 2015), all of which some literature suggest may perform well on tasks involving small corpora (Hamilton et al., 2016a; Altszyler et al., 2017; Oortwijn et al., 2021). Other than models specifically targeting smaller corpora, more recent SemEval-style shared tasks in Russian and Spanish (RuShiftEval (Kutuzov and Pivovarova, 2021) and LSCDiscovery Zamora-Reina et al. (2022)) have shown that Word-in-Context (WiC) and Word Sense Disambiguation (WSD) models tend to have quite high performance on the task of semantic change detection. The WiC models "DeepMistake" (Arefyev et al., 2021; Agarwal and Nenkova, 2022) or XL-LEXEME (Cassotti et al., 2023), or the WSD model "GlossReader" (Rachinskiy and Arefyev, 2021, 2022) may be ideal candidates for future evaluation.

5 Results and Evaluation

For each model described in Section 4, we ran experiments to evaluate performance on downsampled datasets (Section 5.1), quantify variability across bootstrap resamples (Section 5.2), and analyze performance across corpora size (Section 5.3). All reported results are Spearman's rank-order correlation coefficient ρ between the predicted and gold-standard lists of target words ranked by degree of semantic change, as is standard across the literature (Schlechtweg et al., 2020).

5.1 Downsampled Results

We downsampled the SemEval-2020 Task 1 corpora five hundred different times according to the method proposed in Section 3 across all languages and evaluated the models discussed in Section 4 on these downsampled corpora.[7]

We found that Pražák et al.'s model performs the best on average across languages ($\rho = 0.269$), although the gap is small to the model of Rosin and Radinsky (2022) ($\rho = 0.226$), while Pömsl and Lyapin's model, which won the SemEval Subtask 2 shared competition, performs quite poorly, with essentially no correlation demonstrated between predicted and gold standard degree of semantic change lists ($\rho = 0.049$). Interestingly, while Rosin and Radinsky's model performed worse that that of Pražák et al. (2020) when evaluated against the German corpora ($\rho = 0.132$ vs. $\rho = 0.361$), it performed significantly better with the English corpora ($\rho = 0.320$ vs. $\rho = 0.106$). We hypothesize that the difference in performance across these two languages could be due to differing performance in the underlying base models—`bert-tiny` vs. `bert-based-german-cased`—though we leave to future work ablation studies confirming these differences.

Full results are presented in Table 3. On average, there was a 67% decrease in performance compared to the full SemEval corpora, indicating the need for improved methods for detecting semantic change on small corpora.

5.2 Variance Results

In an ideal world, semantic change detection models should display low performance variance: when evaluated on similar datasets they should not have radically different performance. To test whether the models described in Section 4 have this property, we measured the variability in Spearman's ρ across the 500 English downsamples. In these downsamples, only the randomly selected lines change (see

[7] Note that Rosin and Radinsky do not support semantic change detection in Swedish, so we report results only from English and German for their model.

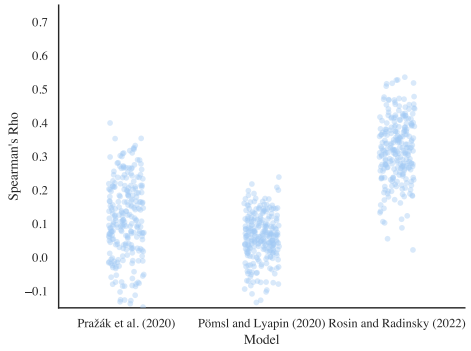

Figure 1: Scatter plot demonstrating the high variance in performance exhibited by each tested model. Each dot represents Spearman's ρ evaluated for a given model on a particular 150k-token downsample of the SemEval English corpus.

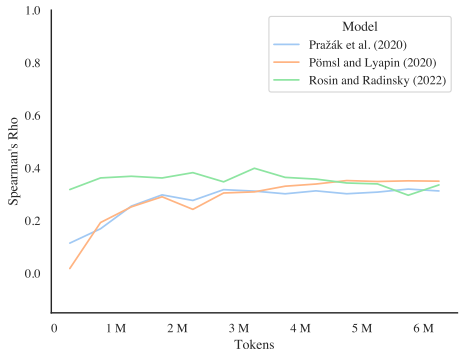

Figure 2: Mean Spearman's ρ across 50 downsamples of the SemEval English corpora plotted against corpus size of both downsampled corpora. The performance of the BERT-based temporal attention model (Rosin and Radinsky, 2022) was essentially stable across corpora sizes, while the performance of the SGNS-based models (Pražák et al., 2020; Pömsl and Lyapin, 2020) improved as corpora size increased.

Table 2) and the gold-standard lines remain the same. Thus, the target words in any two downsampled corpora should present similar degrees of semantic change.

Our results, presented in Figure 1, suggest that all tested models demonstrate startlingly high performance variance. We report the mean performance in the EN column in Table 3, and the standard deviation was: 0.108 for Pražák et al. (2020), 0.075 for Pömsl and Lyapin (2020), and 0.091 for Rosin and Radinsky (2022). This kind of test and result supports the literature studying stability of word embeddings which suggest that small data is especially challenging for the consistency of prediction-based models (Antoniak and Mimno, 2018; Bloem et al., 2019).

5.3 Corpora Size Results

Finally, we evaluated each model across varying sizes of the English corpora. We downsampled the English corpora to individual corpus target token amounts from 250k to 6.25M, with jumps of 500k tokens. We downsampled the corpora 50 times at each token level, with mean Spearman's ρ shown in Figure 2.

For the SGNS-based models of Pömsl and Lyapin (2020) and Pražák et al. (2020), performance improved most dramatically at smaller corpora sizes, although it did generally continue to increase, albeit more slowly, at larger corpora sizes. This was perhaps the expected result, as we believed that more data would improve static embeddings learned from the corpora under consideration. We hypothesize that the reason the performance of the temporal attention model of Rosin and Radinsky (2022) was essentially stable across corpora sizes is due to the author's fine-tuning approach: because the model did not require training from scratch we expected its performance to depend far less on corpus size.[8] These results suggest a model's pre-training could be very influential for semantic change detection performance.

6 Case Study in Literary Intertextual Semantic Change Detection

In the setting of literary criticism, one is often interested in conducting close readings based on subtle differences of language between two books—be they at the level of theoretical motifs, grammatical structures, or single word semiotic shifts—that can then be woven into broader processes of argumentation or productions of comparative meaning (Richards, 1929; Derrida, 1968/2013; Smith, 2016).[9] We propose the application of *literary in-*

[8]The performance of the temporal attention model was not as high as expected nearer to the full SemEval English corpora (at 6.25M tokens mean $\rho = 0.336$ vs. the $\rho = 0.627$ reported in Rosin and Radinsky (2022) on the SemEval English corpora). Despite re-implementing the steps of their paper to the best of our ability, and corresponding with the authors, we were unable to reproduce best reported results from Rosin and Radinsky (2022).

[9]These citations may appear strange to a literary critic, for the study of fluidity in language is embedded in essentially all modern-day literary criticism. We've chosen to cite Richards as *Practical Criticism*'s impact on New Criticism arguably lead to the modern practice of close reading, Derrida because the investigation of slippage in language and

tertextual semantic change detection—the production of a ranked list of words by degree of semantic change between two books—as an exploratory technique to aid literary scholars in finding single word differences that may be of literary interest and suitable for extended investigation (e.g., through comparative close-reading).

As a case study for how, and, importantly to us, whether, current semantic change detection models can be employed to create fruitful avenues of inquiry for literary scholars, we applied the best performing English model evaluated in Section 5.1 (Rosin and Radinsky, 2022) to two books—*The Wretched of the Earth* (Fanon, 1961/2021) and *Scenes of Subjection* (Hartman, 1997/2022). We chose these two novels because we suspected—based on our prior domain knowledge—that they may have interesting intertextual semantic changes (see Section 6.1 for further elaboration). Our research question was: how well does an intrinsic evaluation metric of $\rho = 0.320$ translate to usefulness in an external literary task?

6.1 Case Study Selection

We picked Frantz Fanon's *The Wretched of the Earth* and Saidiya Hartman's *Scenes of Subjection* because we suspected that the word "violence" may have experienced a non-obvious intertextual semantic shift of literary importance. Fanon and Hartman are two black authors writing in a similar literary tradition but whose distinct contexts and research interests shape their interactions with, and study of, violence. To see why this is the case, we will sketch a brief primer of their works.

In 1961, during the Algerian War of Independence, Fanon produced *The Wretched of the Earth*, a searing collection of essays on the psychological effects of colonialism, the effectiveness and cathartic power in violence as a strategy for decolonialization, and the project of post-colonial nation building. Fanon's most radical claims in *The Wretched of the Earth* revolve around his advocacy of physical violence as a productive, beneficial part of decolonialization, a "cleansing force [that] rids the colonized of their inferiority complex, of their passive and despairing attitude [...] emboldens them, and restores their self-confidence" (Fanon, 1961/2021, p. 51).

Hartman, in *Scenes of Subjection*, is interested in a very different kind of violence. She excavates the seemingly small moments of terror and performance that constituted subjection in slavery, what she describes as "the ordinary terror and habitual violence that structured everyday life and inhabited the most mundane and quotidian practices": the ambivalent nature of pleasure mediated in a context of forced performance, the songs enslaved people were made to sing to simulate the appearance of happiness leading up to a coffle, the inability of black bodies to legally bear witness (Hartman, 1997/2022, p. xxx).[10]

We believe, then, that the word "violence" has experienced an intertextual semantic shift suggesting a broader thematic movement of literary significance. If a semantic change detection model can highlight such a shift, then it demonstrates that these systems can be used to suggest avenues of inquiry leading to genuine literary insight. So, our (more specific) research question is: will the model of Rosin and Radinsky (2022) uncover the semantic shift of the word "violence" between *The Wretched of the Earth* and *Scenes of Subjection*? We are also interested in what other terms the model will describe as having experienced semantic shift, and in qualitatively evaluating whether any of those terms have literary importance.

6.2 Literary Validity

Both books were lemmatized and stripped of punctuation. Then non-stopwords that had been used more than 50 times in both books were ranked via the temporal attention model of Rosin and Radinsky (2022) by degree of semantic change.[11] Violence, appearing 367 times across both books, was ranked the tenth most changed word. The top ten words are given in Table 4, as are a small hand-selected series of example sentences we believe suggest the intertextual semantic change that has occurred in the word "violence" between *The*

play in semiotics may have reached its apotheosis with deconstruction and "Plato's Pharmacy", and Smith for her quite lucid article—specifically with a digital humanities audience in mind—on the history and praxis of close reading.

[10] We are essentializing both Hartman's and Fanon's messages for the sake of clarity. Hartman, for instance, is certainly also interested in extreme forms of degradation and violence embedded inside the institutions of slavery, while Fanon was trained as a psychologist and intimately aware of the ways in which colonialism operates as a form of linguistic and cultural violence. Nevertheless, one of Hartman's most impactful contributions was to raise awareness of quotidian forms of violence, and it is difficult to state how impactful Fanon's focus on overt violence remains in academic and radical circles.

[11] For the computational experiment, we used digitized private copies of both books. We cannot make these public due to copyright, but please contact us if interested in reproducibility.

Top-10 words: however, see, since, **political**, new, order, subject, say, life, **violence**

word	Examples from *The Wretched of the Earth*	Examples from *Scenes of Subjection*
violence	the most brutal aggressiveness and impulsive **violence** are channeled, transformed, and spirited away (p. 19).	Songs, jokes, and dance transform wretched conditions into a conspicuous [. . .] display of contentment. This [. . .] itself becomes an exercise of **violence** (p. 53).
	Colonialism is [. . .] naked **violence** and only gives in when confronted with greater **violence** (p. 23).	The most invasive forms of slavery's **violence** lie [. . .] in what we don't see [. . .] mundane [. . .] forms of terror (p. 66).
political	The nationalist **political** parties never insist on the need for confrontation (p. 22).	a notion of the **political** inseparable from [. . .] the ability [. . .] to effect hegemony (p. 109).
	it is not the **political** parties who called for the armed insurrection (p. 32).	What form does the **political** acquire for the enslaved? (p. 109)

Table 4: **Top row:** Top-10 words ranked by degree of intertextual semantic change (greatest first) between *The Wretched of the Earth* and *Scenes of Subjection* according to the temporal attention model of Rosin and Radinsky (2022). **Bottom table:** Hand-selected example sentences demonstrating the semantic change that occurred for "violence" and "political." See Section 6.2 for qualitative interpretation of these semantic shifts.

Wretched of the Earth and *Scenes of Subjection*. Our qualitative evaluation is that in these examples Fanon uses violence to mean *raw physical force producing bodily harm*, while Hartman's use suggests a gentler, though no less injurious, definition: *insidious psychological harm*. As hinted at in footnote 10, these uses are by no means universal throughout the entirety of their respective books, but they do point to what we believe is the broader shift in the way the two authors discuss violence.

Some of the words in the top ten do not seem to have experienced a semantic shift at all. For instance, "however" is ranked the most changed word but seems to be used in a nearly identical and remarkably quotidian way by both Fanon and Hartman. For other words, after conducting close readings of the sentences in which they occur in both novels, we conclude that the shift is unremarkable or mainly an artifact of the distributional hypothesis. "Subject" is a good example. By both Fanon and Hartman, we observe that it is used predominantly to refer to *a person that is discussed, conducted, or investigated*, but Fanon uses it almost exclusively co-occuring with "colonized," as in "colonized subject," while Hartman's more generally uses "subject" to refer to enslaved individuals. Any system based on the distributional hypothesis will determine that "subject" has experienced a semantic shift based on Hartman's lack of use of the term "colonized subject," but it is debatable whether this is an example of semantic shift of literary interest.

More promisingly, the system was able to suggest directions of study which previously we had not considered. "Political," appearing 164 times across both books and ranked the fourth most changed word by the model, is one example of this. Despite having worked with both books extensively, we had not considered "political" as a word or concept with an interesting intertextual semantic difference.

We summarize Fanon's use of "political" primarily in the sense of *in relation to an arm of the administration of the state*, as in "political party," which he uses often. This fits with Fanon's strategic focus, which is at least partly driven by a desire to create a blueprint for actionable political revolution with the aim of divesting unified, anti-democratic political power from colonial governmental regimes. We find that Hartman, in contrast, more often than not uses "political" as a noun signifying *the complex of entanglements existing between a citizen and the state*, as in "the political." Unlike Fanon, there is a somewhat subtle notion in which Hartman questions whether political frameworks are even the right tools through which to understand practices of resistance available to subaltern individuals. She writes that the "traditional notions of the political [. . .] the unencumbered self, the citizen, the self-possessed individual, and the volitional and autonomous subject" are made fraught under slavery, for "Slaves are not consensual and willful actors,

the state is not a vehicle for advancing their claims, they are not citizens, and their status as persons is contested" Hartman (1997/2022, p. 103, 109). The effect of this is that transgressive practices by enslaved individuals—practices of resistance—are made obscure when measured against "traditional notions of the political," for those spheres were not available to and did not encompass slaves. This causes her to both question the suitability of politics as an interpretive device for understanding practices of resistance, and "reimagine the political in toto" Hartman (1997/2022, p. 103, 108). While outside the scope of this work, one could imagine a fruitful investigation and comparative literary paper based on this proposed semantic difference.

6.3 Case Study Discussion

That "violence" was ranked in the top ten most changed words is quite encouraging. For it shows that even with relatively poor performance on the task of determining degree of semantic change in small corpora, as demonstrated in Section 5.1, a semantic change detection system may still produce avenues for investigation that prove viable after sustained literary analysis. Of course, here we already suspected that "violence" had experienced intertextual semantic change. But we did not previously know about the intertextual differences in the word "political." Indeed, "political" is a case study for how we imagine such a system being deployed: take two books, use a semantic change detection system to produce a list of words ranked by intertextual semantic change, and then conduct close readings based on the top ranked words.

We suspect that literary intertextual semantic change detection will be exploratory rather than confirmatory at the ranking stage. One will—and should—always have to return to the text to interrogate whether any suggested word has experienced intertextual semantic change that is both real and of literary interest. We also suspect that to an extent a literary critic must be discerning in order to find words that have interesting intertextual semantic changes. Of the top ten ranked words given in Table 4, only four—political, order, subject, and violence—strike us as being suitable for literary analysis, and some, such as "however," do not seem to us to have experienced any intertextual semantic change at all. It is difficult to know whether this is a product of the relatively poor intrinsic evaluation metrics ($\rho = 0.320$) demonstrated through Section 5.1, a challenge of models based on distributional semantics that have limited ability to understand important surrounding context (be it because of a fixed context window or breaking at the sentence level), or simply one of the impediments of studying unigrams which cannot capture the full spectrum of contextual meaning that literary critics are most interested in studying. Finally, it was extremely labor intensive to determine for each word in the ranked list: a) what the semantic change was, and b) whether the semantic change had literary relevance. This interpretability challenge is perhaps a weakness in existing methods, and an opportunity for future work specifically designed to provide more interpretable output for use in cultural analytics.

We hope as novel methods are developed and intrinsic performance is improved on Experiment 5.1, extrinsic performance on real-world tasks such as this one will become easier and more impactful. Regardless, our case study provided evidence that current semantic change detection systems—even with low intrinsic performance on small corpora—may unveil avenues of investigation in small corpora yielding genuine literary insight.

7 Conclusion

In this paper, we presented—to our knowledge, at least—the first evaluation of semantic change detection models on small corpora (approximately 150k tokens). We found that several high-performing semantic change detection models perform significantly worse on standard tasks evaluated on these smaller corpora, on average experiencing a 67% decrease in performance, and demonstrate remarkably high variance across bootstrap resamples. Overall, for those in the digital humanities there is a clear need for novel and stable methods that are able to accurately detect lexical semantic changes between small corpora, and we hope that our evaluation framework encourages focus on this low-resource setting. However, through a novel literary application and case study, we also demonstrated that there is reason to be optimistic that semantic change detection can be used in an exploratory manner to aid literary critics.

References

plane, n.5. In *OED Online*. Oxford University Press.

Oshin Agarwal and Ani Nenkova. 2022. Temporal ef-

fects on pre-trained models for language processing tasks. *Transactions of the Association for Computational Linguistics*, 10:904–921.

Edgar Altszyler, Sidarta Ribeiro, Mariano Sigman, and Diego Fernández Slezak. 2017. The interpretation of dream meaning: Resolving ambiguity using latent semantic analysis in a small corpus of text. *Consciousness and Cognition*, 56:178–187.

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Nikolay Arefyev, Daniil Homskiy, Maksim Fedoseev, Adis Davletov, Vitaly Protasov, and Alexander Panchenko. 2021. Deepmistake: Which senses are hard to distinguish for a wordincontext model. In *Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*.

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

Arianna Betti, Martin Reynaert, Thijs Ossenkoppele, Yvette Oortwijn, Andrew Salway, and Jelke Bloem. 2020. Expert concept-modeling ground truth construction for word embeddings evaluation in concept-focused domains. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6690–6702, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jelke Bloem, Antske Fokkens, and Aurélie Herbelot. 2019. Evaluating the consistency of word embeddings from small data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 132–141.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234.

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.

Paul Cook and Suzanne Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Jacques Derrida. 1968/2013. Plato's pharmacy. In *Dissemination*. Bloomsbury, London.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy. 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.

Valentina Elce, Giacomo Handjaras, and Giulio Bernardi. 2021. The language of dreams: Application of linguistics-based approaches for the automated analysis of dream experiences. *Clocks & Sleep*, 3(3):495–514.

Martin Paul Eve. 2022. *The Digital Humanities and Literary Studies*. Oxford University PressOxford.

Frantz Fanon. 1961/2021. *The Wretched of the Earth*, 60th anniversary edition edition. Grove Press, New York.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

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

Matthew K Gold. 2012. *Debates in the digital humanities*. U of Minnesota Press.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK. Association for Computational Linguistics.

Thomas Haider and Steffen Eger. 2019. Semantic change and emerging tropes in a large corpus of new high german poetry. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 216–222.

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.

Saidiya Hartman. 1997/2022. *Scenes of Subjection: Terror, Slavery, and Self-Making in Nineteenth-Century America*, 25th anniversary edition edition. Oxford University Press, New York, USA.

Aurélie Herbelot and Marco Baroni. 2017. High-risk learning: acquiring new word vectors from tiny data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 304–309, Copenhagen, Denmark. Association for Computational Linguistics.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Dynamic contextualized word embeddings. 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 6970–6984, Online. Association for Computational Linguistics.

Vani K, Simone Mellace, and Alessandro Antonucci. 2020. Temporal embeddings and transformer models for narrative text understanding. In *Proceedings of the Text2Story'20 Workshop*.

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.

Matthew G Kirschenbaum. 2016. What is digital humanities and what's it doing in english departments? In *Defining digital humanities*, pages 211–220. Routledge.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

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, 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.

Andrey Kutuzov and Lidia Pivovarova. 2021. Rushifteval: a shared task on semantic shift detection for russian. *Computational linguistics and intellectual technologies:*.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

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, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR'13)*.

Stefano Montanelli and Francesco Periti. 2023. A survey on contextualised semantic shift detection. *arXiv preprint arXiv:2304.01666*.

Syrielle Montariol and Alexandre Allauzen. 2019. Empirical study of diachronic word embeddings for scarce data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 795–803, Varna, Bulgaria. INCOMA Ltd.

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.

Kristoffer Laigaard Nielbo, Mads Linnet Perner, Christian Larsen, Jonas Nielsen, and Ditte Laursen. 2019. Automated compositional change detection in saxo grammaticus' gesta danorum. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, Copenhagen, Denmark, March 5-8, 2019*, volume 2364 of *CEUR Workshop Proceedings*, pages 320–332. CEUR-WS.org.

Yvette Oortwijn, Jelke Bloem, Pia Sommerauer, Francois Meyer, Wei Zhou, and Antske Fokkens. 2021. Challenging distributional models with a conceptual network of philosophical terms. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2511–2522, Online. Association for Computational Linguistics.

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

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.

Ondřej Pražák, Pavel Přibáň, Stephen Taylor, and Jakub Sido. 2020. UWB at SemEval-2020 task 1: Lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 246–254, Barcelona (online). International Committee for Computational Linguistics.

Maxim Rachinskiy and Nikolay Arefyev. 2021. Zeroshot crosslingual transfer of a gloss language model for semantic change detection. In *Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*, volume 20, pages 578–586.

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.

I. A. Richards. 1929. *Practical Criticism*. Harcourt Brace & Company, New York City.

Pedro L. Rodriguez, Arthur Spirling, and Brandon M. Stewart. 2023. Embedding regression: Models for context-specific description and inference. *American Political Science Review*, pages 1–20.

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 474–484, New Orleans, Louisiana. Association for Computational Linguistics.

Guy Rosin and Kira Radinsky. 2022. Temporal attention for language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1498–1508.

Guy D. Rosin, Ido Guy, and Kira Radinsky. 2022. Time masking for temporal language models. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 833–841, New York, NY, USA. Association for Computing Machinery.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111, Athens, Greece. Association for Computational Linguistics.

Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. 2021. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135.

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.

Barbara Herrnstein Smith. 2016. What was "close reading"? a century of method in literary studies. *The Minnesota Review*, 2016(87):57–75.

Sandeep Soni, Lauren F. Klein, and Jacob Eisenstein. 2021. Abolitionist networks: Modeling language change in nineteenth-century activist newspapers. *Journal of Cultural Analytics*, 6(1).

Simon Levis Sullam, Giorgia Minello, Rocco Tripodi, and Massimo Warglien. 2022. Representation of jews and anti-jewish bias in 19th century french public discourse: Distant and close reading. *Frontiers in Big Data*, 4.

Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors. 2021. *Computational approaches to semantic change*. Number 6 in Language Variation. Language Science Press, Berlin.

Nina Tahmasebi and Thomas Risse. 2017. On the uses of word sense change for research in the digital humanities. In *Research and Advanced Technology for Digital Libraries*, pages 246–257, Cham. Springer International Publishing.

Rohit Voleti, Julie M. Liss, and Visar Berisha. 2020. A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):282–298.

Claudia Volpetti, K. Vani, and Alessandro Antonucci. 2020. Temporal word embeddings for narrative understanding. In *Proceedings of the 2020 12th International Conference on Machine Learning and Computing*, ICMLC 2020, pages 68–72, New York, NY, USA. Association for Computing Machinery.

Yang Xu and Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *CogSci*.

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 Pre-processing and Cleaning for Cross-Reference

To cross-reference the context text of each annotated use with its SemEval-2020 Task 1 counterpart, we first pre-processed and cleaned the datasets. This was required prior to exact matching due to formatting differences that make straightforward comparisons—and fuzzy matching—inaccurate. We should note that this cleaning procedure was used only to cross-reference. Once a cleaned line from one of the SemEval corpora was matched with the cleaned context text for an annotated use, we inserted the unaltered SemEval line into our downsampled corpora to preserve the properties of the original dataset.

For instance, the English SemEval line:

> period of its greatest activity be towards the middle of the day the hour at which student generally which unfortunate class be most obnoxious to its attack_nn – be unwilling to be disturb.

corresponds to the annotated use:

> period of its greatest activity be towards the middle of the day , the hour at which student generally , - - which unfortunate class be most obnoxious to its attack , – be unwilling to be disturb .

This is a relatively simple example, where stripping punctuation, part of speech tags, and spaces would allow an exact match to be used. However, there are other instances where OCR artifacts[12] or inconsistent formatting made the cross-referencing task slightly more difficult. For example, there was inconsistent formatting in German corpora dealing with the letter "x" in the context of an example like"2x4" (sometimes it is removed, sometimes it is not). To clean our data for cross-referencing, then, we stripped punctuation, OCR artifacts, duplicate and trailing spaces, _nn and _vb part of speech tags, and finally the letter "x" from both the lemmatized context text for each annotated use and each line from the lemmatized SemEval corpora.

We found only one exception that couldn't be cross-referenced with this procedure and manually included it in the final dataset. The SemEval line:

> so after the famous christmas-dinner with its nice roast-meats and pudding and pie after the game of romp with her father and the ride on the rocking-horse with her brother who at last from mere mischief have tip_vb her off and send her cry to her mother begin to think about go there

corresponds to the following context text surrounding the word "tip":

> so , after the famous christmas-dinner with its nice roast-meats , and pudding , and pie , - - after the game of romp with her father , and the ride on the rocking-horse with her brother , who , at last , from mere mischief , have tip her off , and send her cry to her mother , —she begin to think about go there .

The discerning reader will notice that there is one word missing ("her mother begin to think" vs. "her mother , —she begin to think") in the SemEval corpora.

[12]We provide examples of these OCR artifacts in our code repository https://github.com/jnehrenworth/small-corpora-scd.

We attempted to develop similar heuristics for the Latin dataset, but we were unable to do so because of larger formatting and content inconsistencies between Latin context text and SemEval lines. For more detailed documentation, visit `downsample.py` of our repository: https://github.com/jnehrenworth/small-corpora-scd.