Sense-specific Historical Word Usage Generation

Pierluigi Cassotti
University of Gothenburg, Sweden
pierluigi.cassotti@gu.se

Nina Tahmasebi
University of Gothenburg, Sweden
nina.tahmasebi@gu.se

Abstract

Large-scale sense-annotated corpora are important for a range of tasks but are hard to come by. Dictionaries that record and describe the vocabulary of a language often offer a small set of real-world example sentences for each sense of a word. However, on their own, these sentences are too few to be used as diachronic sense-annotated corpora. We propose a targeted strategy for training and evaluating generative models producing historically and semantically accurate word usages given any word, sense definition, and year triple. Our results demonstrate that fine-tuned models can generate usages with the same properties as real-world example sentences from a reference dictionary. Thus the generated usages will be suitable for training and testing computational models where large-scale sense-annotated corpora are needed but currently unavailable.

1 Introduction

Language is essential to almost every aspect of human life and is often a crucial remnant of previous societies. It is through previous writings that we can learn how societies were built and have since evolved. But as our societies evolve, so does our language. Great effort has been made to support historical linguistics, for example, by using computational approaches to model how words change their meaning over time, a phenomenon called semantic change. However, these efforts have been severely hindered by a lack of diachronic sense-annotated corpora, primarily because sense-annotation of historical texts requires significant time, expertise, and effort and is thus extremely costly (Schlechtweg et al., 2024).

Our contribution: We have brought together the generative abilities of large language models (LLMs) and the vast resources of a high-quality, large-scale dictionary to train and evaluate generative models using the real-world example sentences provided with each sense in the Oxford English Dictionary (OED). Specifically, we:

1. introduce fine-tuned LLMs capable of generating large and realistic diachronic, sense-annotated datasets (Section 4);
2. provide a comprehensive suite of evaluation tools (Section 5); and
3. show that synthetic generated usages can be used to effectively train models (Section 7).

These resources open up the possibility of developing models for a range of different tasks, ranging from diachronic word sense disambiguation to semantic change detection.

Using generative models, we can produce sentences in which a word is used in a semantically and historically accurate way. Given a *word* and a *sense definition*, these models can generate temporally accurate sentences for each specified time period. In Table 1, for different specified years (1980–2010), we see that the *same* definition of *phone* results in sentences using phone cradle, phone booth, and cell phone in an accurate way that historically represents the word *phone*.

We show that fine-tuned LLMs (1) can generate sentences that have the same properties as the original example sentences from the OED with a time error of (on average) 50 years; and (2) can be used to simulate synthetically large-scale linguistic phenomena such as semantic change, providing accurate and diverse historical word usages. We provide a suite of tools for testing (a) context variability, (b) temporal accuracy, and (c) semantic accuracy of generated historical usages. We have also released a new human-annotated dataset of semantic relatedness between generated historical sentences and sense definitions.[1] The annotations show empirical evidence of a correlation between

[1]The fine-tuned models and the tools for evaluation, including a sentence-dating model and a WSD (Word Sense Disambiguation) model, are available in the HuggingFace Hub. The annotated usage-definition pairs are published in Zendodo. The code is available in Github.

Transactions of the Association for Computational Linguistics, vol. 13, pp. 690–708, 2025. https://doi.org/10.1162/tacl_a_00761
Action Editor: Alessandro Lenci. Submission batch: 10/2024; Revision batch: 2/2025; Published 7/2025.

Year	Usage
1980	He put the **phone** back in the cradle and turned toward the kitchen.
1990	We made the telephone call from a public telephone booth that had **phones**, instruments and paper on top of each other in a jumble.
2000	I was carrying my cell **phone** so that I could hear the signals if there were any.
2010	You can buy a low-priced 3G **phone** (the phone's the thing, not the service) for less than £60. It's no phone you'd be proud to have on your desk, but it'll handle text messaging, e-mail and internet-browsing.

Table 1: Generated usages for **phone** with the same definition '*A telephone apparatus; a telephone receiver or handset*' over time.

the hierarchical organization of senses in the OED and human perceptions of the closeness between the word senses.

2 Motivation

Large diachronic sense-annotated corpora are needed to develop models for longitudinal studies. Such corpora could be obtained through dictionaries that store implicitly sense-annotated word usages via example sentences attached to each sense. However, these usages are sparse, even in large dictionaries. For example, in the OED each word sense is accompanied by approximately four sentences, which are spread across an average of four distinct years and sometimes even fewer. For example, the *corruption* sense of the word *graft* appears only in a single year.

When explicitly sense-annotated datasets are available, they are often small and synchronic (Miller et al., 1993; Pasini and Camacho-Collados, 2020). To simulate diachronic datasets, they can be divided into batches, but they then lack historically accurate language. To retain a historically accurate linguistic style, efforts have been made to derive sense information implicitly by clustering similarity judgments between pairs of usages of a word (Schlechtweg et al., 2021), as these are cheaper to produce than sense annotation. However, these are also small in scale and typically only cover two time periods.

To train and evaluate models for longitudinal studies, for example, to detect semantic change over time, we need sense-specific and temporally changing word usages on a large scale, that is, with sufficient representation of each sense across different historical contexts. Example sentences or small-scale sense-annotated data are insufficient for this purpose. However, using fine-tuned LLMs, we can generate the desired number of sentences that match the historical context *and* specific senses of words for training and test models. We can thus ensure that the generated output has a continuous and representative temporal distribution of any and all word senses across all (for the sense's valid) time periods.

3 Data

In this work, we used the Oxford English Dictionary (OED, 1989) because it is the most comprehensive and authoritative record of the English language. It includes over 273,000 words with detailed etymologies and (frequency) bands.[2] For each word, there is a set of entries that contain a group of senses that are semantically related (see Figure 1). Multiple entries for a word reflect either different parts of speech or homonymic senses. For each sense, a set of example sentences offer the reader a sense of how a word can and has been used. A sense entry is introduced when there is evidence of the sense in writing, and a sense is considered rare or outdated when it is no longer in everyday use.

We collected words, senses, and examples using the OED Research API (OED API) from the year 1700 onward. The dataset was split into two distinct partitions, one for training and the other for testing. Table 2 shows the statistics for the training set and the test set. The training set included all the available parts-of-speech (PoS) entries in the OED, while the test set contained the four PoS tags that are most susceptible to semantic change: nouns, verbs, adjectives, and adverbs. For the test set, we considered, for each word, only the senses with at least five examples of usage in the OED. We further aimed to choose a set of words that together were as comprehensive as possible in terms of the range of polysemy and frequency, as well as in terms of the representation of senses over time.

The estimation of word polysemy was based on the number of main senses reported in the OED, appearing on the second level or above in the hierarchy (in yellow in Figure 1). The lemmas were classified into three categories: (1) those

[2]From Google Books Ngrams (v2). Due to a lack of sense annotation, no sense-level frequency is available.

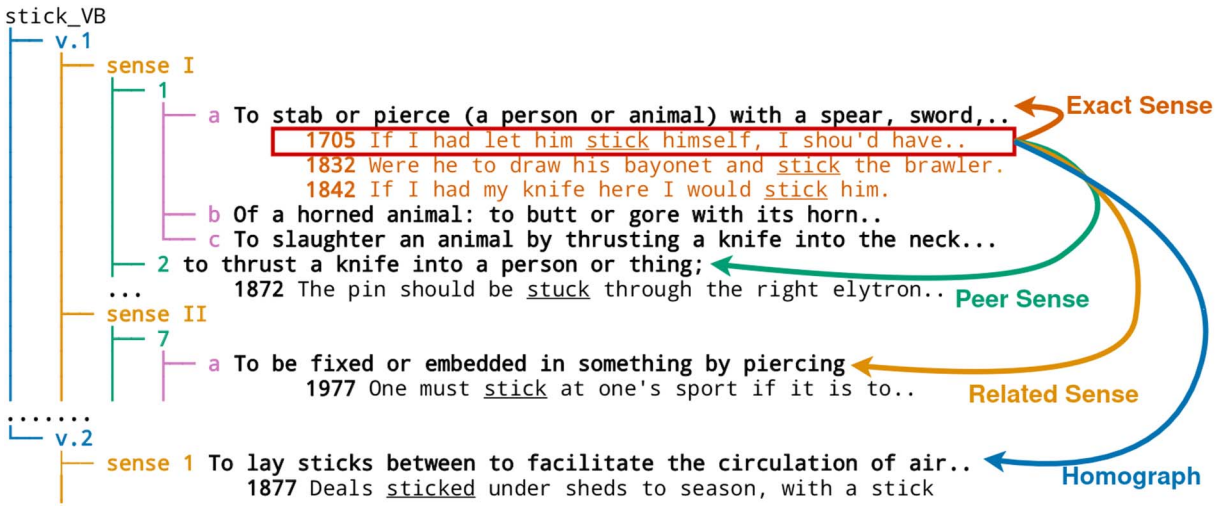


Figure 1: The OED hierarchical representation of the senses of the verb *stick* and the relation of the example sentence (highlighted by a red rectangle) to the respective sense definition (`Exact Sense`) and other sense definitions, including `Related Sense`, `Peer Sense`, and `Homograph`.

		Train	Test
Words	**Lemmas**	81,220	2,768
	Part-of-speech	11	4
Senses	**Senses**	301,395	13,762
	Avg. definition length	128.01	146.70
	Avg. number of usages per sense	3.95	6.52
Usages	**Usages**	1,191,851	89,759
	Avg. text length	91.22	90.29

Table 2: OED statistics. The large difference in the avg number of usages per sense stems from the requirement of at least 5 example sentences when choosing test words.

with at least one new sense introduced after 1800, (2) those with a disappeared sense, and (3) all remaining lemmas. Stratified sampling was used to ensure a uniform distribution across the various classes of word frequencies, polysemy, and introduction/extension of senses. The training and test data were entirely distinct from one another; the lemmas that appear in the test data and their respective senses were removed from the training data.

4 Generation of Usages

The primary focus of our work is a proof-of-concept using an open source model, for which we chose Llama. We fine-tuned the pretrained Llama 3 8B model using the OED dataset, producing two versions: Llama 3-Janus and Llama 3-Janus-PoS.[3] Janus was trained without PoS data, while Janus (PoS) incorporated PoS tags. Then we tested fine-tuned models against the instructed models. Specifically, we used the instructed Llama 3 8B Instruct, and Llama 3 70B Instruct, consisting, respectively, of 8 and 70 billion parameters (AI@Meta, 2024).

We also undertook a comparison to GPT as a commercial tool, working with GPT-3.5 and GPT-4o as the most efficient and effective models provided by OpenAI (OpenAI).

4.1 Notation

A large language model is denoted by $\mathcal{M}$, a target word by w, and the respective sense definitions by d_i for $i = 1, 2, \ldots, n$. For each target word and sense d_i the OED attests a validity period (t_i, t_j) such that the sense d_i is recorded for the word w. We denote y as a year in (t_i, t_j). We query the model $\mathcal{M}$ with triples $(w, d_i, y) \in \mathcal{T}$ such that it produces a usage u of a word w in the sense d_i with language appropriate for year y (see Table 1 for examples of usages generated for a triple). The set of usages is denoted $U_{w,d_i,y}$. For the OED, a usage is an example sentence, and y is the year in which the sentence originated. For generated sentences, we can choose y as any and all years in the validity period (t_i, t_j). To reduce the amount

[3]From now on, these models will be referred to as *Janus* and *Janus (PoS)*, respectively. In Roman mythology Janus is the god of transition and time. He is depicted with two faces, symbolizing the ability to see both the past and the future.

of evaluation, we sampled a single year y for the set of triples $\mathcal{T}$.

4.2 Prompting Instructed Models

We prompted each of the instructed models $\mathcal{M}$, namely GPT-3.5, GPT-4o, Llama 3 8B Instruct, and Llama 3 70 B Instruct. We used the same prompt P for each of them, i.e., $\mathcal{M}(P, w, d_i, y)$ for all $(w, d_i, y) \in \mathcal{T}$. Additionally, for GPT-4o we tested the few-shot strategy (Few shot GPT-4o) providing 5 illustrative examples of the task. Further details of the prompting, including the prompt itself, are given in Appendix B.

4.3 Fine-tuning

We fine-tuned the pretrained Llama 3 8B model on the 1,191,851 historical sense-annotated usages extracted from the OED, which constitute the training set described in Section 3. Specifically, the fine-tuning process uses causal language modeling (CLM), where the model is trained to predict the usage tokens, given the input sequence consisting of a concatenation of the target year y, the target lemma w, and the target sense definition d_i delimited by the special token <|t|>, i.e.,

$$y <|\mathrm{t}|> w <|\mathrm{t}|> d_i <|\mathrm{s}|> u <|\mathrm{end}|>$$

For Janus (PoS), we provide additionally the word part of speech p:

$$y <|\mathrm{t}|> w <|\mathrm{t}|> d_i <|\mathrm{p}|> p <|\mathrm{p}|><|\mathrm{s}|> u <|\mathrm{end}|>$$

The designated special token $<|\mathrm{s}|>$ indicates the beginning of the completion segment, i.e., the usage field, which the model will predict in an autoregressive manner, ensuring that the model uses the context provided by the year, lemma, and definition to accurately generate the subsequent usage example. To reduce the training effort and reduce costs, we used quantization and low-ranking adaptation (QLORA) (Dettmers et al., 2023). Further details of the implementation are given in Appendix C.

5 Evaluating Generated Usages

In this section, we present the evaluation of the usages generated by Janus, Janus (PoS), Llama 3 8B Instruct, Llama 3 70B Instruct, GPT-3.5, GPT-4o, and Few Shot GPT-4o. This evaluation is based on the OED test set, as introduced in Section 3. For each entry in the OED test set, we added a triple with a sampled year (w, d_i, y) to $\mathcal{T}$. Using each model $\mathcal{M}$, we then generated ten usages for all triples in $\mathcal{T}$, resulting in $U_{(w,d_i,y)}$ containing a total of 137,620 generated usages for each model. For each $\mathcal{M}$, we evaluated the set of generated usages $U_{(w,d_i,y)}$ focusing on:

- **context variability** λ: how diverse the contexts are within the set of generated word usages, evaluated as $\lambda(U_{(w,d_i,y)})$;
- **semantic accuracy** σ: how accurately each of the generated word usages conveys the target definition d_i, measured as $\sigma(u, d_i)$;
- **temporal accuracy** τ: how well each of the generated word usages u is historically coherent with the target year y, $\tau(u, y)$. For instance, the word *airplane* should never appear in contexts before 1903.

Validation: To validate our results (and obtain a baseline for λ, τ and σ), we compared the generated usages with the original usages included in the OED test set. Furthermore, we validated context variability and temporal accuracy, using a **control dataset** without sense annotations comprising usages for each $(w, *, y) \in \mathcal{T}$. The control set thus consisted of historical word usages $U_{(w,*,y)}$ extracted from the Corpus of Late Modern English texts (De Smet, 2005) (1710–1920) and the Clean Corpus of Historical American English (Alatrash et al., 2020) (1800–2010).

5.1 Context Variability

A key goal was that the generated usages should have diverse contexts to penalize models that paraphrased the same sentence over and over again (an example of this can be found in Table 11 in the Appendix). For each set of generated usages $U_{(w,d_i,y)}$, we measured the variability in three different ways:

- **Lexical Overlap:** The similarity between generated usages was measured using:
  - (**Jaccard**) The ratio of shared words between two usages to the total number of unique words[4]
  - (**BLEU**) An extension of Jaccard that considers word sequences (n-grams), accounts for repetition, and applies a brevity penalty to discourage overly short usages

[4]We tokenize text using whitespace and remove stopwords based on the NLTK stopword list `https://www.nltk.org/`.

Source	Usage	Definition	Label
Janus	No sooner had we parted than he called up and said that he had made a decision 'to make things **stick**' in the West.	Of a plan, order, decision, etc.: to be complied with or implemented; to be permanently effective. Frequently in to make (something) stick.	4 (Exact Sense)
		To lay sticks between (timber boards) in order to facilitate the circulation of air during seasoning.	1 (Homograph)
OED	He is the man of all others slow to admit the thought of revolution; but let him once admit it, he will carry it through and make it **stick**.	transitive. To pierce (something) with a sharp-pointed object; to prick, puncture. Frequently with specifying the sharp-pointed object.	2 (Related Sense)
		To be reluctant or unwilling (to do something); to hesitate, to scruple. Chiefly in negative constructions (e.g., he did not stick to)	3 (Peer Sense)

Table 3: Examples of both positive and negative triples derived from Janus-generated usages and the OED test set. The examples illustrate different semantic relations, including Exact Sense, Peer Sense, Related Sense, and Homograph. Each row shows a usage example, its corresponding sense definition, and the label indicating the semantic relation.

- **Paraphrase Index (Cosine):** the similarity of each usage pair was computed as the cosine between the usage vectors encoded using SBERT[5] (Reimers and Gurevych, 2019)
- **Semantic Similarity (BScore):** the BERTScore[6] (Zhang et al., 2020) was used to evaluate semantic similarity of usage pairs.

The final score for each $\lambda_j \in \{$Jaccard, BLEU, Cosine, BScore$\}$ over the set of usages was calculated as the average of the individual usage pair (u, v) scores, given by $\frac{1}{|Q|}\sum_{(u,v)\in Q}\lambda_j(u,v)$, where $Q = \{(u,v) \mid u, v \in U_{w,d_i,y}, u \neq v\}$.

For the control dataset, which had no information on word sense, we aggregated at the lemma level, i.e., $U_{(w,*,y)}$. For the OED test set, which does not have enough usages for each year considered, we aggregated all years and evaluated the diversity of usages for the specific sense, i.e., $U_{(w,d_i,*)}$. When comparing the control and the generated usages, we should therefore bear in mind that both validation datasets naturally exhibit a greater variety of contexts than the generated uses. The usages of the control dataset show a greater variety along the sense dimension, whereas the usages of the OED test set show a greater variety along the temporal dimension.

The results in Table 4 highlight the context variability of the usages generated by different models compared to the control dataset and the OED test set. For example, the latter has a Jaccard index of 0.03, a cosine similarity of 0.31, and a BScore of 0.85. Notably, Janus and Janus (PoS) stand out with the lowest BLEU score (0.21/0.20) of the generated data sets, only slightly less diverse than the control data set (0.18) and the OED test set (0.19). This result indicates a higher context variability among the generated usages of fine-tuned models compared to the usages generated by instructed models, where the generated sentences are more similar to each other. For example, GPT-3.5 shows a cosine similarity score of 0.58 and a high BScore of 0.90, suggesting that there is a lot of repetition across the usages. The usages generated by the Llama 3-8B and Llama 3-70B Instruct models exhibit moderate variation. Because of their broader and more varied contexts, the control dataset and the OED test set show the expected high variability.

5.2 Semantic Accuracy

In this section, we want to assess the semantic accuracy of the generated uses, i.e., how well the meaning of each generated usage $u \in U_{(w,d_i,y)}$ reflects the provided definition d_i (Erk et al., 2013). To assess the semantic accuracy of a generated usage, we want the usage to reflect the meaning of d_i but not other related meanings $d_j,\ j \neq i$. For each usage $u \in U_{(w,d_i,y)}$ in the OED test set, we created up to three *negative usages* by sampling definitions d_j from peer senses, related senses, and homographs, respectively (see an example in Table 3). The generated usages were then evaluated both computationally and by human annotators. We measured performance using the

[5]all-mpnet-base-v2.

[6]roberta-large.

Dataset	Jaccard	BLEU	Cosine	BScore
GPT-3.5	0.16	0.40	0.58	0.90
GPT-4o	0.13	0.38	0.58	0.90
Few Shot GPT-4o	0.19	0.42	0.59	0.91
Llama 3-8B Instruct	0.08	0.29	0.47	0.88
Llama 3-70B Instruct	0.17	0.36	0.51	0.90
Janus	0.04	0.21	0.35	0.86
Janus (PoS)	**0.03**	**0.20**	**0.32**	**0.85**
Control dataset	0.08	0.18	0.23	0.83
OED Test set	0.03	0.19	0.31	0.85

Table 4: Context variability. Labels from Section 5.1. Validation datasets in *italic*. We expect low values for all four measures, as the generated sentences should be diverse.

Dataset	Spear. Correlation	F1 score
GPT-3.5	**0.64**	**0.77**
GPT-4o	**0.64**	0.76
Few Shot GPT-4o	**0.64**	**0.77**
Llama 3-8B Instruct	0.62	0.76
Llama 3-70B Instruct	0.63	0.76
Janus	0.61	0.75
Janus (PoS)	0.59	0.74
OED Test set	0.76	0.83

Table 5: Semantic accuracy for each model generated dataset. Validation dataset in *italic*.

macro F1 score and Spearman's correlation. The macro F1 score is calculated for two classes: Exact Sense and Different Meaning (Peer Sense, Related Sense, Homograph). Model predictions (or human annotation scores) below 3 were categorized as Different Meaning. Spearman's correlation measures the agreement between the predictions of the WSD (word sense disambiguation) regression model and human annotations or OED labels. The results indicate that, while automatic evaluations show that all models performed similarly in capturing semantic relations (Section 5.2.1), human evaluations reveal that Janus-generated usages were semantically comparable to the original OED usages (Section 5.2.2).

5.2.1 Computational Evaluation

Unlike traditional WSD methods that predict sense labels, our approach mimics human annotation by comparing usage-definition pairs, allowing for direct comparison of human and model performance. We trained a WSD classifier on the OED training data. In particular, given a definition d_i and a usage u, the classification model predicts the semantic relations extracted from the OED, namely, Exact Sense (4), Peer Sense (3), Related Sense (2), and Homograph (1). Further details on the classification model are reported in Appendix E.

The results in Table 5, while showcasing the overall strong performance of the models in capturing semantic relations, also reveal challenges in the evaluation process. The results show that all models, including Janus, exhibited strong Spearman's correlation (ranging from 0.59 to 0.64) and high F1 score (ranging from 0.74 to 0.77). However, the gap between these scores and that of the OED test set (0.76/0.83), which serves as a reference point, indicates that while the models demonstrate proficiency in capturing semantic relations, there remains room for improvement.

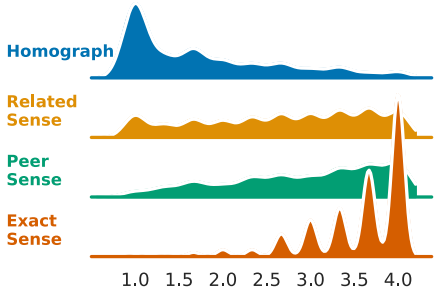
One issue relates to the labels extracted from the OED (Peer Sense, Related Sense, and Homograph), which are sometimes inconsistent. This inconsistency stems from the *lumpers and splitters* issue in lexicography, where lexicographers differ in the way they group senses together or split them into finer distinctions, e.g., what could have been an exact sense was instead labeled as peer sense. This subjectivity can introduce noise, potentially lowering model performance.

5.2.2 Human Evaluation

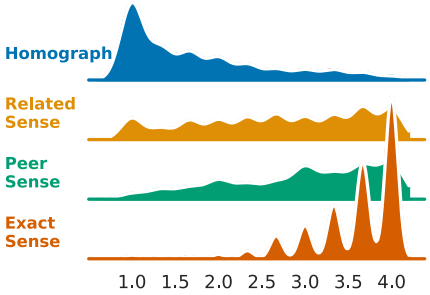
The dataset employed for the automatic evaluation consists of 259,489 and 1,043,311 usage-definition pairs for the OED test set and Janus, respectively. We therefore simultaneously subsampled a set of 2584 usage-definition pairs both from the Janus and the OED test sets. To make the dataset as diverse as possible we ensured that there was only one usage for each sense and only one positive triple for each word. We assessed semantic accuracy by manually annotating whether the word meaning in the Janus-generated usages matched the definition provided. To validate the results, we used the OED test usages and assessed semantic accuracy in a corresponding way. We asked annotators to determine the semantic relatedness of the definition with respect to the provided usage, using the DURel semantic relatedness scale proposed by Schlechtweg et al. (2018). In particular, as shown in Table 6, we assumed a match between the DURel scale and the organization of the meaning in the OED. Each usage-definition pair

↑	4: Identical	Identity	Exact Sense
	3: Closely related	Context variance	Peer Sense
	2: Distantly related	Polysemy	Related Sense
	1: Unrelated	Homonymy	Homograph

Table 6: The relation between the DURel scale (left), the respective continuum of semantic proximity (middle) (Blank, 1997), and the OED derived labels (right).



(a) OED Test set



(b) Janus

Figure 2: Ridge plot showing the distribution of OED labels across increasing human annotated scores. The x-axis represents the human annotated scores, while the y-axis represents the density of different OED labels.

was annotated by three annotators, and the final score was obtained by averaging their ratings.

We used Amazon Mechanical Turk Amazon Mechanical Turk for evaluation. More information on the annotation guidelines, the cost, and the annotator agreement statistics are reported in Appendix F. Figure 2 shows the result of the evaluation and the distribution of the DURel scale scores across sense labels. For example, when word usages with completely unrelated senses (i.e., homonyms) are paired, the majority of the humans label the relation as 1. When evaluated against human annotations, both the Janus-generated usages and the original OED usages exhibit comparable performance in terms of Spearman's correlation and F1 score (Table 7). This suggests that from a human perspective,

Dataset	Spear. Correlation	F1 score
Janus	0.57	0.72
OED Test set	0.58	0.72

Table 7: Annotation results. Spearman's Correlation and F1 score are calculated for the OED labels compared to the human annotations, taking into account both Janus and OED usages.

the OED usages and the Janus usages are effectively indistinguishable by non-experts in terms of their semantic accuracy. Similarly, in both cases, related sense and peer sense were barely distinguishable.

5.3 Temporal Accuracy

We next turn to temporal accuracy, referring to whether the generated usage matches the year given during generation. Evaluating this is typically very complex and challenging for humans, as distinguishing between texts from different decades, such as the 1950s as opposed to the 1990s, requires considerable expertise. Because such expertise is difficult to find among available annotators, we automated this task and broadened the classification from a year to the decade in which the year occurs.

However, measuring temporal accuracy is also a computationally challenging task. Although considerable work has been done in the past, the focus has often been on larger segments of text, such as paragraphs or entire documents (Vashishth et al., 2018; Kanhabua and Nørvåg, 2009). It is generally accepted that classifying a single sentence, as opposed to a paragraph or a full document, is more difficult due to the limited context (Popescu and Strapparava, 2015).

We fine-tuned roberta-large using the usages from the OED training set to classify the decade in which a given usage was written. The classification spans 33 decades from 1700 to 2020. Further details on the decade classifier can be found in Appendix D.

Table 8 shows results with low accuracy (0.13–0.15 on the validation datasets), highlighting the complexity of correctly classifying the *exact* decade. These results are in line with what has been demonstrated in previous work. The root mean squared error (RMSE) results are more meaningful. The RSME results on the *control dataset* show that the decade classification model

Dataset	RMSE	Accuracy
GPT-3.5	143.44	0.05
GPT-4o	147.52	0.03
Few Shot GPT-4o	125.38	0.05
Llama3-8B Instruct	129.92	0.06
Llama3-70B Instruct	108.09	0.07
Janus	54.75	**0.12**
Janus (PoS)	**52.69**	**0.12**
Control dataset	47.97	0.13
OED Test set	52.75	0.15

Table 8: Temporal accuracy. RMSE and Accuracy are reported for each model. Validation datasets in *italics*, generated datasets above.

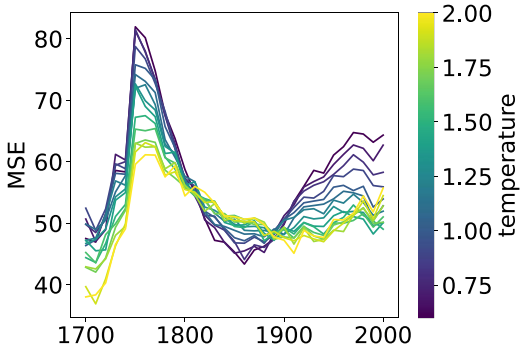


Figure 3: Root mean square error over time for Janus using different temperatures.

is valid as it predicts the decade of sentences with an error of just 47.97 years. In this dataset, while classification may be easier due to longer sentences offering more context, the model is challenged by having to classify sentences that are filled with OCR errors, tokenized and extracted from different resources. For the OED test set on the other hand, the sentences are typically shorter and we get a slightly higher RMSE of 52.75.

Fine-tuned Janus models proved able to generate usages that were comparable to the validation datasets, with an RMSE of (54.75/52.69), similar to that of the OED test set (52.75). All other models had an RMSE at least twice that of the Janus models, with the Llama 3 70B Instruct slightly ahead of 8B Instruct.

The RMSE of Janus for different temperature settings is shown in Figure 3. The values were very low around the year 1700, where it is likely that the capitalization of nouns allows the classifier to correctly assess the decade, but spike about 50 years later. This spike is probably due to the challenges of generating texts from that era, which are masked by the prevalent noun capitalization of the early 18th century. After the year 1800, the values normalize and we find a temperature effect: The higher the temperature, the more temporally accurate the generated sentences. This could stem from the training data being dominantly modern. To access less-likely phrasing and wording, a higher temperature setting is needed to allow the model to access long-tail knowledge.

6 In-depth Analysis

Semantic Accuaracy and Word Types We conducted a detailed analysis of semantic accuracy based on the scores obtained from the WSD model, across factors such as part of speech, word frequencies, and outdated senses. We measured the average predicted scores within the same sense where we expected perfect scores close to 4. To assess the statistical significance of the mean differences among different groups, we used ANOVA tests, all of which yielded statistically significant results (p-value < 0.01). In terms of parts of speech, the average predicted scores were lower for adverbs (average score of 3.70) compared to nouns (3.78), verbs (3.80), and adjectives (3.84) indicating difficulty recognizing the same sense of an adjective across different usages. Word frequency also played a role: More frequent words (bands 1–5) achieved higher scores (3.85) than less frequent words (bands 6–14), which scored an average of 3.78. Additionally, the ability to identify usages of the same outdated sense–those considered archaic or obsolete–was lower (average score of 3.37) compared to non-outdated senses (3.80).

However, when we examined the human evaluation data, the differences were not statistically significant for any of the studied dimensions. This may be due to the fact that (i) in the computational evaluation, the difference we observe between the groups is due to a bias in the WSD model used for the evaluation rather than in Janus; or (ii) the annotated sample size is insufficient to derive significant differences.

Temporal Effect on Semantic Accuracy We found that the temporal dimension significantly impacts the annotation process. In previous work, including our own, it has been assumed that the nature of annotation (WiC-style) is so general that extensive knowledge of the historical period is unnecessary. As a result, employing annotators

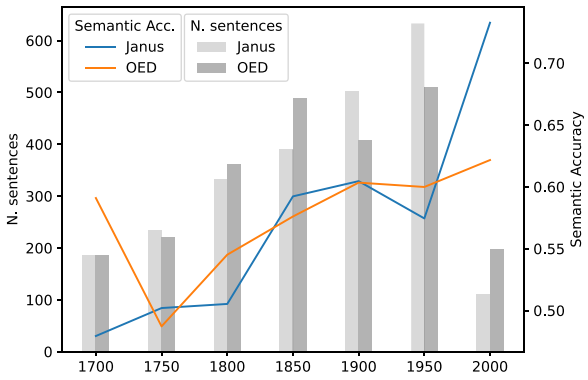


Figure 4: Semantic accuracy over time (human annotation). The bar chart represents the number of sentences evaluated in each time interval (left axis), while the line chart shows the semantic accuracy correlation for Janus and OED dataset (right axis) across historical periods.

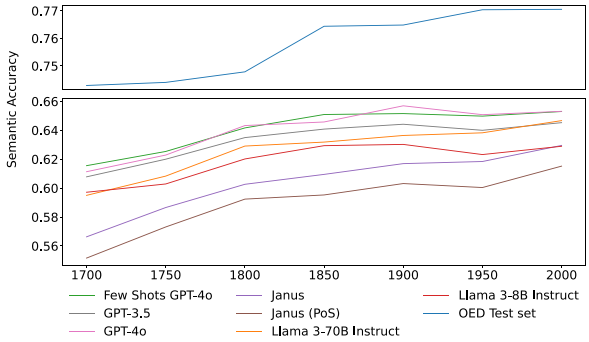


Figure 5: Semantic accuracy over time (computational evaluation). Correlation over time, computed between WSD model scores and OED scores. OED test set (top), generated usages (bottom).

with specific expertise in history or historical linguistics has not been considered essential for this task. However, in Figure 4, we report semantic accuracy (Spearman correlation) calculated based on the year each sentence was written, with results grouped into 50-year intervals. Semantic accuracy improves over time, rising from 0.45 to over 0.7. Examples generated by Janus and those from the OED exhibit a similar trend, both reflecting this upward progression. For completeness, Table 13 in Appendix F also reports the agreement between annotators, calculated based on examples from different time periods. Notably, the period from 1900 to 2000 shows the highest level of agreement among annotators.

The historical period also influences computational evaluation. In Figure 5, we show the correlation of Semantic Accuracy, calculated by the WSD model, based on the usages generated

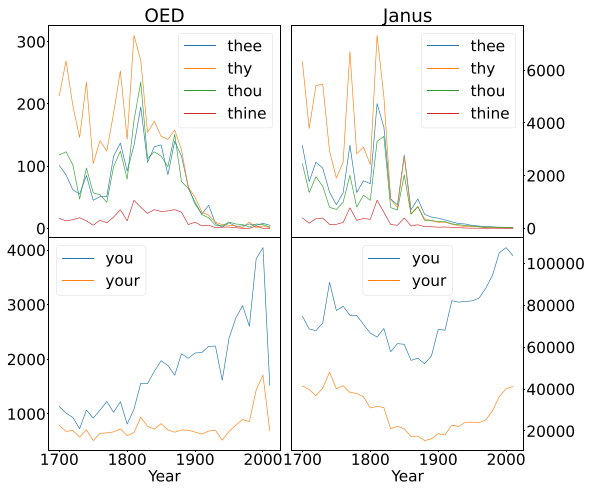


Figure 6: Temporal word choice. The figure shows archaic pronouns (thee, thy, thou, thine) declining sharply after 1800 and nearly disappearing by the late 19th century, while modern pronouns (you, your) rise steadily from the 18th century.

by the models and the OED test set. The Semantic Accuracy for the WSD model increases over time, particularly up until 1850, after which it appears to stabilize. In general, we observe the same trend for all models, indicating that fine-tuned models behave similarly to instructed models over time while achieving higher temporal accuracy. The decrease in performance in earlier periods cannot be attributed to stylistic changes, as noted in Section 5.3, where we observed that instructed models exhibit a gap of over 100 years. Instead, it is more likely due to a greater prevalence of archaic word senses in historical texts, which models trained primarily on contemporary data struggle to handle.

Temporal Analysis of Language and Semantic Shifts in the Janus-generated Usages We analyzed how Janus uses temporal cues to generate historically accurate texts. We generated a diachronic corpus from 1700–2010 with 100 usages per decade for each entry in the OED test set. We focused on context words (excluding the target word w) to assess the fine-tuned models implicit understanding of language across decades. For example, we examined the frequency of archaic pronouns like *thee*, *thy*, *thou*, and *thine* (Figure 6) and found a decreasing frequency, while *you / r* had an increasing frequency. We explored Janus's awareness of temporal shifts in word meaning (Figure 7) by noting that it uses a context word

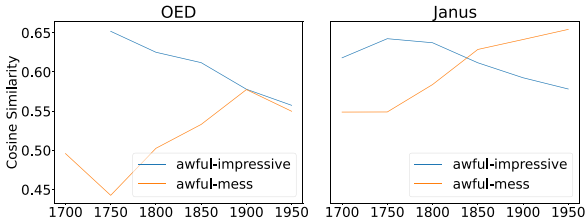


Figure 7: Temporal sense-awareness. Cosine similarity trends between *awful* / *mess* and *awful* / *impressive* computed using XL-LEXEME (Cassotti et al., 2023) embeddings, show a semantic shift. Initially linked to *impressive*, *awful* began associating with *mess* around 1850, similar to what was found in Hamilton et al. (2016).

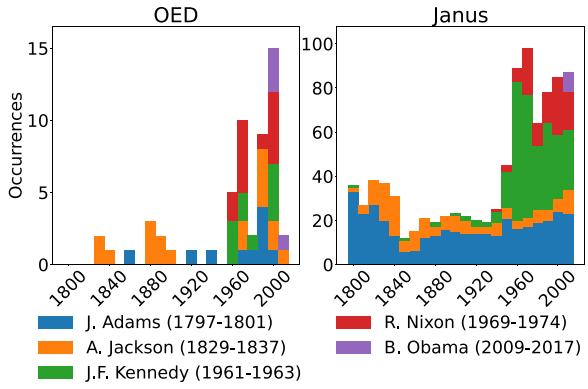


Figure 8: Temporal relevance. The plot shows mentions of US presidents over time: Adams is consistently mentioned from 1790, Jackson gains prominence from 1820, and Obama only appears after 2010, reflecting their historical timelines.

like *awful* in a temporally accurate way. Additionally, we evaluated how Janus handles relevant named entities over time, such as US presidents (Figure 8). We found that Janus correctly references presidents in their appropriate historical contexts, mentioning Obama only after 2010 and Nixon after 1960.

Out-of-distribution Usage Generation Building on studies like Periti et al. (2024), which highlight the challenges of LLMs with unseen meanings, we explored Janus's ability to generate usages beyond its training distribution. We analyzed seven cases where Janus was tested on fictional or historically inaccurate word senses (Table 10 in the Appendix). When there is a strong semantic link to existing senses, in, e.g., generalization or auto-antonym shifts, Janus relies more heavily on pretrained knowledge and ignores the definition provided by generating ambiguous sentences (w = train, d_i = 'A conveyance, a form of transport.' u = *I'm a huge fan of the* ***train*** *and believe it is the future of public transport in London.*) or using negation in the case of auto-anonym shift (w = good, d_i = 'Of poor quality or little worth.' u = *I hope your year isn't* ***good****, I hope it's great, I hope you go a long way and see things you've never seen.*). When we instead introduce a domain shift, e.g., metaphorical extension, the sentences are nicely aligned with the definition provided (w = zebra, d_i = 'A mixture of dissimilar qualities or traits.' u = *The* ***zebra*** *of his playing: a big and powerful stroke on one side, and a delicate tap on the other*).

7 Application of Generated Usages

We aim to verify whether automatically generated word usages can be used effectively for training and evaluating models in downstream applications. Specifically, we choose the Lexical Semantic Change Detection (LSCD), where the aim is to determine how much a word has changed over time based on usages from different time periods. The LSCD pipeline typically involves, for each word, sampling of usages from two different time periods. Next, these usages are compared pairwise (either from within each time period, or across periods) and assigned a score using the DuREL scale (1–4, Table 6). Considering these usages as nodes and the scores as edges, a Diachronic Word Usage Graph (DWUG) is constructed and the usages are clustered using graph clustering. Change scores are obtained by analyzing (frequency) changes of the resulting clusters.

For this study, we focus on the first step, the pairwise comparison of usages, commonly known as graded Word-In-Context (WiC) task. We first evaluate using the standard English LSCD dataset, the DWUG EN dataset (Schlechtweg et al., 2022a), which consists of 46 words with annotated usage graphs with sentences sampled from time periods 1810–1860 and 1910–1960. Next, for the same 46 words, we take example sentences from the OED, WordNet (Miller, 1992) and SemCor (Miller et al., 1993) and construct WiC datasets. These are constructed such that example sentences from the same sense are labeled as 1 and from difference senses are labeled as 0. The example sentences from OED are taken from two historical periods (1800–1900 and 1900–2000), while WordNet+SemCor have only modern sentences. Using these datasets, we want

Model	WordNet+SemCor	OED		DWUG EN		
		1800–1900	1900–2000	1810–1860	1910–1960	All
roberta-large	0.37	0.32	0.39	0.36	0.45	0.38
+ WN	0.36	0.51	0.63	0.41	0.52	0.46
+ OED (1800)	0.46	0.62	0.68	**0.48**	0.55	0.50
+ OED (2000)	**0.49**	0.62	**0.73**	**0.48**	**0.59**	**0.53**
+ OED (ALL)	0.46	**0.65**	**0.73**	**0.48**	0.58	0.52

Table 9: Spearman correlation for roberta-large and its fine-tuned versions on the Word-in-Context (WiC) task. The fine-tuned models are trained on synthetic usages generated by Janus using definitions from WordNet (WN) and the Oxford English Dictionary (OED). The models are evaluated on real-world usages from WordNet+SemCor, OED, and DWUG EN.

to know if Janus generated, synthetic usages can help improve model performance on the graded word-in-context task.

To fine-tune the models, we generate Janus usages for the 46 words of DWUG EN using definitions from both WordNet (WN) and the Oxford English Dictionary (OED). Since WordNet is a synchronic dataset, we generate examples for the year 2000. For the OED, we generate usages corresponding to the years 1800 and 2000, as well as a combined set (OED ALL) that includes examples from both time periods. For each word-definition pair in these datasets, we generate 100 usages.

For model training, we adopt a Siamese encoder approach following Cassotti et al. (2023). The target word is highlighted using special tokens (`<t>` and `</t>`) surrounding it. Each usage pair is encoded with *roberta-large*, and the centroid of the subword vectors is used as the final usage representation. The model is trained with a contrastive loss to ensure that vectors of word usages with the same meaning are closer in space, while those with different meanings are farther apart.

Table 9 presents the performance of different models: *roberta-large* and its fine-tuned versions using generated usages (+WN, +OED). The baseline *roberta-large* model achieves moderate performance, with scores ranging between 0.32 and 0.45, performing best on DWUG EN (1910–1960). Fine-tuning *roberta-large* on synthetic usages derived from small datasets leads to notable improvements, despite the limited training data (only 46 words from DWUGs). *roberta-large*+WN exhibits stronger performance on OED (1800–1900) and OED (1900–2000), while *roberta-large*+OED achieves the highest scores, particularly in OED (1900–2000) and DWUG EN (1910–1960).

8 Related Work

Generating Dictionary Examples Barba et al. (2021) represents the first attempt to employ generative models for producing novel word usage examples for previously unseen dictionary entries. Specifically, this work fine-tunes BART (Lewis et al., 2020), an encoder-decoder model, and the evaluation process focuses on: (i) estimating the semantic accuracy of the generated sentences by classifying them using a WSD model based on BERT (Devlin et al., 2018), and (ii) assessing fluency (whether the sentence is logical and grammatically correct) and coherence (whether the word's meaning aligns with its given definition). He and Yiu (2022) introduce the first constrained generation approach for example generation. The constraints are designed to generate sentences that optimally reflect the reader's educational level, offering control over both readability and sentence length. Cai et al. (2024) evaluate example generation by prompting various LLMs. Their evaluation combines automatic methods with human validation, assessing whether the generated sentences are preferred over real-world examples. The results indicate a preference for the generated sentences.

These previous studies relied on a smaller subset of the OED dictionary without temporal information (Gadetsky et al., 2018). Our work goes beyond previous work by generating example sentences over long timespans while taking the temporal language into account. For this reason, we created a new dataset using the OED API that preserves the year from which the example sentences stem. Our dataset differs from previous ones in both size (it is significantly larger) and OED version (using a later version). Unfortunately, the absence

of sense identifiers in the previous dataset makes direct comparison challenging.

Our work is the first to assess both temporal accuracy and context diversity–two fundamental aspects in LSCD applications. Furthermore, unlike previous studies, our evaluation of semantic accuracy extends beyond ensuring that the sentence aligns with the provided definition. We also focus on ensuring that the sentence is unambiguous and does not express other meanings. Therefore, our method penalizes cases where the usage does not distinctly reflect the intended meaning. For example, the term *meat* used to refer broadly to food but has narrowed to mean flesh. The OED provides clear examples of this shift, such as ''First take all the *meat* out of the lobster,'' where *meat* clearly refers to flesh. However, in broader contexts like ''He had *meat* and drink enough,'' the meaning of *meat* becomes more ambiguous. By accounting for the word's other meanings, we make our evaluation more robust and better suited for testing the creation of semantic change datasets.

Resources for LSCD While semantic change can be modeled on any temporal dataset, prior to 2020 evaluation relied on small datasets of words with attested meaning changes in lexicographic resources (Hamilton et al., 2016; Rudolph et al., 2016; Yao et al., 2018; Frermann and Lapata, 2016). SemEval-2020 Task 1 (Schlechtweg et al., 2020) marks the first systematic approach to evaluation grounded in the texts themselves using the DURel framework (Schlechtweg et al., 2018).

Despite this progress with follow-ups in several languages, two key limitations remain: (1) explicit word sense labels, and (2) longitudinal datasets covering multiple time periods. The former can be addressed by splitting *synchronic* sense-annotated corpora (Schlechtweg and Schulte im Walde, 2020), but inherits the limitation of both the small size and the synchronic nature of the corpus. The latter is addressed using *diachronic* corpora and the introduction of synthetic changes by altering word frequencies or contexts using a replacement schema (Kulkarni et al., 2015; Shoemark et al., 2020; Dubossarsky et al., 2019). Although this produces a naturally changing linguistic style, the sentences are not natural (e.g., the *chair* was purring loudly).

Generative Models for LSCD When modeling semantic change, LLMs are usually used in one of two ways (Periti and Montanelli, 2024): (1) *as computational annotators*, where instruct-based models have been used to predict the annotation between sentence pairs (Karjus, 2023; Wang and Choi, 2023; Periti and Tahmasebi, 2024); or (2) to generate definitions for each usage of a word, either using prompting or fine-tuning (Giulianelli et al., 2023; Fedorova et al., 2024). Despite these advances, there is a notable gap in the literature regarding the ability of these models to augment data for historical texts. Although LLMs have access to vast amounts of data, historical data is underrepresented compared to modern data, as the majority of their training data is sourced from the Web. This paper shows that LLMs specifically fine-tuned to model the temporal dimension can generate text that is temporally accurate to a substantial degree.

9 Conclusion

In this paper, we have demonstrated significant progress in the generation of sense and time-specific text using LLMs. We have shown that Llama, when fine-tuned, can generate historically accurate text and that higher temperature settings improve stylistic authenticity at the expense of semantic accuracy—a consequence of the long-tail distribution problem. We have successfully developed a model capable of generating historical texts with properties that closely mirror example sentences found in the OED. Qualitative investigation of the generated sentences shows that in the context of a target word w, the model uses temporally accurate clues such as word choices (e.g., thee, thou, you); and correct senses of the context words (e.g., gay used as happy, awful as impressive).

Our contributions include the development of a time classifier, a sense classifier, and a robust methodology that can be utilized in future research and adapted for new models of historical text generation. Additionally, we provide a human-annotated dataset of word usages paired with the corresponding sense definitions, a valuable resource for further studies. Our empirical evidence confirms a strong alignment between the DURel scale and the hierarchical structure of the OED, reinforcing the validity of our approach. Janus opens new avenues for the study of lexical semantic change by enabling model training and testing with accurate historical

sense-annotated data over extensive time periods, where previously, no such data existed.

Our findings demonstrate that the temporal dimension influences both model performance and annotation outcomes. Future research should consider these factors by more thoroughly investigating the role of temperature settings in models and by employing experienced annotators. The ambiguity of the OED labels, influenced by the lumpers and splitters issue, presents intriguing avenues for research. Moreover, while our work currently focuses on English, there is potential to extend it to other languages by testing Janus's zero-shot capabilities or by fine-tuning on additional languages.

Acknowledgments

This work has in part been funded by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021). The computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

We would also like to express our gratitude to Asad Basheer Sayeed, Francesco Periti, Dominik Schlechtweg, as well as the TACL reviewers and action editor for their valuable feedback on the preliminary draft of this work.

References

AI@Meta. 2024. Llama 3 model card.

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean corpus of historical American English. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020*, pages 6958–6966, Marseille, France. European Language Resources Association.

Amazon Mechanical Turk. [link].

Edoardo Barba, Luigi Procopio, Caterina Lacerra, Tommaso Pasini, and Roberto Navigli. 2021. Exemplification modeling: Can you give me an example, please? In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19–27 August 2021*, pages 3779–3785. ijcai.org. https://doi.org/10.24963/ijcai.2021/520

Andreas Blank. 1997. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*, volume 285 of *Beihefte zur Zeitschrift für romanische Philologie*, Niemeyer, Tübingen. https://doi.org/10.1515/9783110931600

Bill Cai, Ng Clarence, Daniel Liang, and Shelvia Hotama. 2024. Low-cost generation and evaluation of dictionary example sentences. 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 3538–3549, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.194

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. https://doi.org/10.18653/v1/2023.acl-short.135

Hendrik De Smet. 2005. A corpus of late modern english texts. *ICAME Journal*, 29(2005):69–82. https://doi.org/10.20885/informatika.vol3.iss1.art7

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10–16, 2023*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28–August 2, 2019, Volume 1: Long Papers, pages 457–470. https://doi.org/10.18653/v1/P19-1044

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554. https://doi.org/10.1162/COLI_a_00142

Mariia Fedorova, Andrey Kutuzov, and Yves Scherrer. 2024. Definition generation for lexical semantic change detection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5712–5724, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.339

Lea Frermann and Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45. https://doi.org/10.1162/tacl_a_00081

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry P. Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, Volume 2: Short Papers*, pages 266–271. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2043

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.176

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 2116. ACL. https://doi.org/10.18653/v1/D16-1229, PubMed: 28580459

Xingwei He and Siu-Ming Yiu. 2022. Controllable dictionary example generation: Generating example sentences for specific targeted audiences. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22–27, 2022*, pages 610–627. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.46

Nattiya Kanhabua and Kjetil Nørvåg. 2009. Using temporal language models for document dating. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2009, Bled, Slovenia, September 7–11, 2009, Proceedings, Part II*, volume 5782 of *Lecture Notes in Computer Science*, pages 738–741. Springer. https://doi.org/10.1007/978-3-642-04174-7_53

Andres Karjus. 2023. Machine-assisted mixed methods: Augmenting humanities and social sciences with artificial intelligence. *arXiv preprint arXiv:2309.14379*.

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. https://doi.org/10.1145/2736277.2741627

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Barbara McGillivray, Dominik Schlechtweg, Haim Dubossarsky, Nina Tahmasebi, and

Simon Hengchen. 2021. Dwug la: Diachronic word usage graphs for latin. https://doi.org/10.5281/zenodo.5255228

George A. Miller. 1992. WORDNET: A lexical database for english. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, USA, February 23–26, 1992*. Morgan Kaufmann. https://doi.org/10.3115/1075527.1075662

George A. Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. 1993. A semantic concordance. In *Human Language Technology: Proc. of a Workshop Held at Plainsboro, New Jersey, USA, March 21–24, 1993*. Morgan Kaufmann. https://doi.org/10.3115/1075671.1075742

Oxford English Dictionary OED. 1989. Oxford english dictionary. *Simpson, Ja & Weiner, Esc*, 3.

OED API. [link].

OpenAI. [link].

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

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. https://doi.org/10.18653/v1/2024.acl-long.246

Francesco Periti and Stefano Montanelli, New York, NY, USA. 2024. Lexical semantic change through large language models: A survey. *ACM Computing Surveys*, 56(11). https://doi.org/10.1145/3672393

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. https://doi.org/10.18653/v1/2024.naacl-long.240

Octavian Popescu and Carlo Strapparava. 2015. Semeval 2015, task 7: Diachronic text evaluation. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4–5, 2015*, pages 870–878. The Association for Computer Linguistics. https://doi.org/10.18653/v1/S15-2147

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Maja R. Rudolph, Francisco J. R. Ruiz, Stephan Mandt, and David M. Blei. 2016. Exponential family embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 478–486.

Dominik Schlechtweg, Pierluigi Cassotti, Bill Noble, David Alfter, Sabine Schulte Im Walde, and Nina Tahmasebi. 2024. More DWUGs: Extending and evaluating word usage graph datasets in multiple languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14379–14393, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.796

Dominik Schlechtweg, Haim Dubossarsky, Simon Hengchen, Barbara McGillivray, and Nina Tahmasebi. 2022a. Dwug en: Diachronic word usage graphs for english. https://doi.org/10.5281/zenodo.7387261

Dominik Schlechtweg and Sabine Schulte im Walde. 2020. Simulating lexical semantic change from sense-annotated data. In *The Evolution of Language: Proceedings of the 13th International Conference (EvoLang13)*.

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, NAACL-HLT, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2027

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, SemEval@COLING2020*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.1

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2022b. Dwug de: Diachronic word usage graphs for German. https://doi.org/10.5281/zenodo.7441645

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 *Annual Conference of the North American Chapter of the Association for Computational Linguistics, (NAACL-HLT 2021)*. Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.567

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2020. Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1007

Nina Tahmasebi, Simon Hengchen, Dominik Schlechtweg, Barbara McGillivray, and Haim Dubossarsky. 2022. Dwug sv: Diachronic word usage graphs for swedish. https://doi.org/10.5281/zenodo.7389506

Shikhar Vashishth, Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha P. Talukdar. 2018. Dating documents using graph convolution networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, Volume 1: Long Papers*, pages 1605–1615. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1149

Ruiyu Wang and Matthew Choi. 2023. Large language models on lexical semantic change detection: An evaluation. *arXiv preprint arXiv:2312.06002*.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018*, pages 673–681, Marina Del Rey, CA, USA. ACM. https://doi.org/10.1145/3159652.3159703

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020*. OpenReview.net.

A OED Test Set Statistics

The dataset consists of 2,768 lemmas, which include 357 adverbs, 760 verbs, 773 adjectives, and 994 nouns. Within this dataset, there are a total of 13,762 senses, of which 6,607 are identified as main senses, 2,697 were introduced after 1800, and 355 have become outdated. In Figure 9, we present the distribution of words according to their frequency band and the number of main senses. The highest concentration of words (as indicated by the yellow color) appears in the region where the frequency band is around 10 and the number of main senses is around 4. As we move from left to right on the x-axis (increasing frequency band), there is a general trend of increasing word count

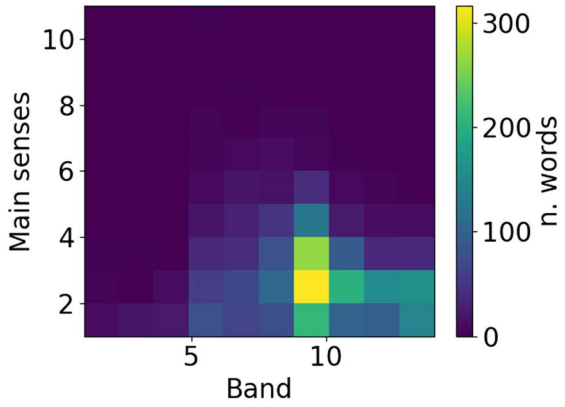


Figure 9: Heatmap showing the distribution of words based on their frequency band (x-axis) and the number of main senses (y-axis) in the dataset. The color intensity indicates the number of words within each category, with yellow representing the highest concentration.

until around band 10, after which it diminishes. Similarly, as we move up the y-axis (increasing number of main senses), the number of words seems to initially increase but then decreases, indicating that while some words have multiple senses, it is less common for very frequent words to have many senses.

B Prompting

We used the `n` parameter of the OpenAI API's GPT-3.5 and the `num_return_sequences` parameter of the HuggingFace API to generate multiple responses to the same prompt, creating different examples for the same definition, lemma, and year. The prompt is:

Retrieve a sentence from the year **`year`** *where the word* **`lemma`** *means* **`definition`**. *If you can't find one, create a fictional sentence. The answer should only contain the sentence.*

Here, **`year`**, **`lemma`**, and **`definition`** are replaced with the target values from the OED. We consistently used the system prompt: *You are a helpful assistant*. For Few-shots GPT-4o, we provided five input-output pairs as examples[7] to guide the model in generating relevant sentences. We post-processed the output from instruct-based models using regular expressions, looking for sentences within quotation marks or following a specific pattern, like text after a colon. Once a sentence was found, it was cleaned by removing extra characters such as newlines and quotes. We then verified that the target word appears in the cleaned sentence, noting its position.

[7]The full prompt is provided in the Github repository.

C Fine-tuning

We fine-tuned the Llama3-8B pretrained model using QLORA with 4-bit precision (nf4 quantization) and half-precision (float16) computations, so reducing memory usage and enabling training on less powerful hardware. The tokenizer was customized with special tokens <|s|>, <|t|>, <|end|>, and [PAD], and <|p|> for Janus (PoS) to denote the part of speech tag. Sequences were left-padded for consistency. Input data were formatted with specific markers for sections such as year, lemma, and definition, and tokenized to a maximum length of 512 tokens. Training was configured with a batch size of 4, gradient accumulation over 2 steps, and a learning rate of 0.0002, running for 1 epoch with checkpoints saved at the end of each epoch. The LORA configuration includes an alpha set to 16, a dropout of 0.1, and a rank of 8.

D Decade Classification

We fine-tuned the `roberta-large` model for sequence classification with 33 labels. The dataset was tokenized using a tokenizer from the base model, with texts truncated and padded to a set length. The data were split into 90% for training and 10% for validation. We trained the model for one epoch using a learning rate of 1e-6 and a batch size of 32. Evaluation and model saving occurred at every 10% of the total steps, with performance measured by the weighted F1 score after each evaluation.

E Sense Classification

We fine-tuned roberta-large for a regression task, predicting the OED labels (Homograph (1), Peer Sense (2), Related Sense (3), or Exact Sense (4)). The model was trained using the MSE loss, treating the task as a continuous prediction problem. The tokenizer was extended with two special tokens (<t> and </t>) to mark the target word in the input text. The model's token embeddings were resized to include these tokens. The dataset was split 90-10 into training and validation sets, and the text was tokenized with truncation and padding. We trained the model for one epoch with a learning rate of 2e-5, a batch size of 32, and a weight decay of 0.01.

Case	Year	Lemma	Definition	Generated Usage
Auto-antonym shift (Not attested)	2020	good	Of poor quality or little worth.	I really did not like the new **good** quality product on the car. What does your mother think of this **good** looking man? I hope your year isn't **good**, I hope it's great, I hope you go a long way and see things you've never seen.
Auto-antonym shift (Not-attested Period)	1800	awful	Used to emphasize something unpleasant or negative.	He was so angry at his son's conduct that he threatened an **awful** blow.
Auto-antonym shift (Not-attested Period)	2020	awful	Demonstrating or indicative of profound reverence or respect.	The ceremony begins with the **awful** entry into the chamber of a priestess.
Neologism	2020	tably	Of a surface: Without curvature, indentation, or protuberance; plane, level.	He was on **tably** level ground.
Metaphorical shift (Not attested)	2020	zebra	A mixture of dissimilar qualities or traits.	The **zebra** of his playing: a big and powerful stroke on one side, and a delicate tap on the other.
Generalization shift (Not attested)	2020	train	A conveyance, a form of transport.	He'd rather drive a four-wheel-drive, six-speed, automatic **train** with a big engine and lots of body. I'm a huge fan of the **train** and believe it is the future of public transport in London.
Specialization shift (Not-attested Period)	1800	record	The known history of the life or career of a person, esp. a public figure.	He is one of the most amiable, gentle and benevolent men that ever lived and, in the **record** of his life ...

Table 10: Generated usages for different cases of pretrained knowledge testing.

F Annotation

We used Amazon Mechanical Turk to collect annotations from crowd workers based in the United States, Canada, the United Kingdom, or Australia. Workers spent an average of 16 seconds per sentence pair, earning 3.5 cents per pair. The study cost a total of $926.35, with $573.09 going to workers, $168.13 to MTurk fees, and $185.31 to taxes.

The annotators were provided with the following guideline:

Determine how closely the definition matches the meaning of the bolded word in the sentence, using a scale from 1 (Unrelated) to 4 (Identical), or select 'Cannot decide' if you are unsure.

The annotation results, as summarized in Table 12, highlight key differences between the DWUG datasets and our study. Unlike the DWUG EN, SV, and DE datasets, which are annotated word usage pairs, our study focused on sense definition usage pairs, similar to the DWUG LA dataset. However, DWUG LA achieved a higher Krippendorff's alpha (0.62) than ours (0.52). This difference may stem from DWUG's use of highly skilled annotators, such as PhD students, and smaller, carefully selected groups, leading to more consistent annotations. Our larger pool of over 100 annotators introduced more variability, likely reducing overall agreement and impacting Krippendorff's alpha. Additionally, in Table 13 we report the Krippendorff's alpha computed over time.

Model	BScore	Generated Sentences
Few-Shot GPT4-o	0.89	The chimney did **reek** thick smoke into the evening air, much to the dismay of the townsfolk.; The factory chimneys **reek** out thick clouds of smoke, darkening the sky.; Forth from those jovial fair Assemblies, where the gay reeked out full in their Faces, the hot and **reeking** Scents of Wine and Meat.; The chimneys now **reek** out the dense fog that covers the village each evening.; The chimney did **reek** out a thick, black smoke that darkened the sky.; The **reeking** pest moistens the ambient air, while the thick sulphur clouds ascend in curling wreaths.; Factory chimneys, belching and **reeking** out their noxious plumes, darkened the sky above the industrial town.; The chimneys did **reek** out a thick and murky smoke that clouded the village.; War was but a pastime, a sport, **reeking** out blood and fire.; The factory did seem to **reek** out great clouds of noxious smoke, casting a pall over the entire village.
Janus	0.82	I have often remarked how many of the principal towns in England are smokeless and unhealthful, owing to a want of chimnies, and not **reeking** the smoak of their own fires, the houses remain damp and cold.; I was sitting i' my chair, and smoking my pipe with my gudeman, when in came a man **reeking** out his fum.; That the whole of this vapour or gas is not **reekin** up to the top of the mountain is apparent.; The smoke he **reeks** out is far more sweet than that which arises from his own faggots.; I **reekit** the stink of the stoor as I saunter about.; A watter, that a body may haud in the mou and **reeke** in the air for a lang time.; The steam of a cold stove will **reek** more strongly out of a clean than a soiled pan.; The sma', **Reikin** thro' the crumblin' glou' a snaw.; As for the **reek** of Peculiarities I do not see what you are about?; The fire will **reek** more than smoke, if any one of your old family were so poor as to go to begg.

Table 11: Example of context variance. Usages generated for the word *reek* in the year 1769, based on the definition: *transitive. To exhale, emit, or give out (smoke, fume, vapour, etc.). Frequently with out. Frequently figurative and in extended use.*

Dataset	KRI
DWUG LA (McGillivray et al., 2021)	0.62
DWUG EN (Schlechtweg et al., 2022a)	0.61
DWUG SV (Tahmasebi et al., 2022)	0.56
DWUG DE (Schlechtweg et al., 2022b)	0.53
This paper	0.52

Table 12: Annotators' agreement. The Krippendorff's alpha for ordinal data (KRI) is reported for different datasets. Higher is better.

Year	1700	1750	1800	1850	1900	1950	2000
KRI	0.50	0.50	0.45	0.52	0.54	0.53	0.55

Table 13: Annotators' agreement over time.