# Investigating the Laws of Synonymy Evolution through a Fine-Grained Lemma-Based Framework

Anonymous ACL submission

## Abstract

We investigate how synonyms evolve over time, focusing on two apparently contradictory tendencies: the Law of Differentiation, whereby synonyms diverge in meaning, and the Law of Parallel Change, whereby they co-evolve. We develop a framework that breaks down the laws into a fine-grained taxonomy, which we use to construct a diachronic dataset based on the Oxford English Dictionary. We rely on definition embeddings and the use of LLM-as-a-Judge to link synonyms. We validate our methodology by performing a human evaluation on a subset of the constructed dataset. Finally, we populate the taxonomy and conduct extensive analyses across different time spans to investigate the predominance of each law in English over time. We find that while the Law of Differentiation is larger in absolute numbers, it is driven by the general tendency for meanings to get lost rather than by a property unique to synonymy.[1]

## 1 Introduction

Words that express the same meaning or refer to the same thing are considered *synonyms*. Two synonyms need not always have been synonyms, nor is there any guarantee that they will remain so in the future. Some, including Haiman (1980) and Leclercq and Morin (2023), believe that, for pragmatic reasons, lexical differences always correspond to semantic differences, and thus synonymy is an unstable relation in the long run. On the other hand, it has been claimed, both for efficient and expressive needs, that the availability of multiple forms for the same meaning is a design feature of languages (Gardner et al., 2021; Levshina and Lorenz, 2022). On this view, words that share meaning but are used in different registers and situational contexts may remain synonymous.

The way synonymic relationships between words arise and develop over time has long been a central topic in historical linguistics. Already a century ago, linguists put forward hypotheses about recurring patterns in synonym development, including two apparently competing laws. The **Law of Differentiation** (LD; Bréal, 1897; Sturtevant, 1917) states that synonyms tend to diverge over time; while the opposing **Law of Parallel Change** (LPC; Stern, 1921, 1931) states that, once synonymous, words tend to co-evolve. The empirical evaluation of these laws can provide insight into the general tendencies that govern lexical evolution and the long-term dynamics of semantic relations. Due to the manual effort involved, these laws were only evaluated on a small set of synonyms.

Previous computational studies (Xu and Kemp, 2015; Lietard et al., 2023) have attempted to examine these two laws on a larger scale and determine whether one prevails over the other, producing contrasting results. This inconsistency in previous studies stems from two main aspects. First, LD and LPC have so far been operationalized as a binary relationship: two words either diverge or change in parallel. In reality, the process is far more complex: two words may, e.g., simultaneously encounter multiple different changes to their set of meanings. Second, in the absence of reliable lexical resources to ground analyses of the phenomenon, previous studies have relied on computational and empirical models of word meaning. Although lexical and corpus-based approaches are complementary, the lack of temporally granular, sense-structured lexical data meant that prior work could not fully capture the fine-grained semantic dynamics underlying these laws.

We address these critical issues as follows: (i) we develop a cross-linguistically-applicable framework to formalize and operationalize the LD and LPC through a fine-grained taxonomy of nine categories; these decompose the different ways synonymic relationships may evolve over time (§3); (ii) we ground our analysis in a large-scale dataset

[1] The codebase will be publicly released upon acceptance.

derived from the *Oxford English Dictionary* (OED), which provides temporally anchored word senses together with their definitions and periods of attestation (§4.1); (iii) we investigate the challenging task of explicitly aligning senses across synonymous words by employing both computational models and human annotators on the task of cross-word sense alignment (§4.2); (iv) we conduct extensive analyses of the dataset across time spans of different lengths and compare to a control dataset. Our results shed light on which law has more prevalent support in the resource to help clarify the currently inconclusive picture emerging from previous studies (§5). Although the present study focuses on English, the framework is designed to facilitate future extension to other languages.

## 2 Related Work

Recently, there has been a surge in research on historical language change using computational methods, including work on lexical semantic change (Giulianelli et al., 2020; Tahmasebi et al., 2021; Periti and Montanelli, 2024) with approaches to predict change across two or more time periods (Periti and Tahmasebi, 2024). Most work in this area has focused either on detecting semantic change for individual words (e.g., Schlechtweg et al., 2020; Zamora-Reina et al., 2022; Fedorova et al., 2024), or on formulating and testing general laws of semantic change (e.g., Hamilton et al., 2016; Dubossarsky et al., 2017; Uban et al., 2021).

The study of change at the word-pair level (synonym relations in particular) has received less attention in computational work, with early contributions by Turney and Mohammad (2019) and Mohammad (2020). This is a timely gap to fill, particularly because the role of synonymy has become a topic of intense theoretical debate (§1). Recent work suggests that addressing this debate requires taking into account both the polysemy of individual lemmas and the synonymy relations between them (Franco et al., 2022; Cai and De Smet, 2024), a joint modeling challenge that depends on the availability of large-scale, sense-structured resources and reliable alignment methods.

Computational work that has explicitly targeted the LD and LPC (the two laws central to our work) is limited, and has so far lacked the sense-structured resources needed to model polysemy and synonymy jointly. In a corpus-based study, Xu and Kemp (2015) proposed the first computational method aimed at predicting how synonymous pairs in English evolve over time. Their results indicate evidence in favor of the LPC. Their work relied on normalized co-occurrence vectors to approximate meaning, and, most notably, their predictions could not be evaluated against a ground truth given the resources available at the time.

A follow-up study by Lietard et al. (2023) proposed alternative computational approaches based on SGNS-vectors to investigate the same phenomenon. Their results point in the opposite direction, providing evidence in favor of the LD. This study also presents important limitations. The binary classification in the paper ("Syn." vs. "Diff." in their terminology) does not directly reflect the LPC vs. LD dichotomy. Their Syn. class groups together pairs that underwent parallel change with pairs that remained unchanged, the latter being uninformative with respect to the two laws.

In contrast to previous work, we (i) make use of a high-quality lexical resource (i.e., the OED) as the basis of our investigation; (ii) directly address the question of LD and LPC by operationalizing them through a fine-grained taxonomy; and (iii) construct a diachronic dataset using a meaning- and time-annotated lexical resource.

## 3 Framework

**Terminology** A **meaning** (also called a "word sense") is a discrete semantic unit. A **definition** is a textual realization of a meaning (i.e., a dictionary definition). A **form**, e.g., `meat`, is an abstraction representing all inflectional and historical variants (e.g., *meat*, *mete*, *meats*) of a lemma. A **lemma** is a dictionary entry with a part-of-speech (POS), consisting of a form together with a set of meanings, expressed via definitions, and their associated periods of attestation. Two lemmas are **synonyms** if they share at least one meaning. Figure 1 presents an example with the lemmas $\underline{\text{food}}_{(N)}$ and $\underline{\text{meat}}_{(N)}$, and their sets of meanings at two time periods.

**Original Formulations of LD and LPC** The LD is attributed to Bréal (1897, p. 30) and Sturtevant (1917). The former formulated LD as "Synonyms do not exist for long: either they differ, or one of the two terms disappears",[2] while the latter as "Unless synonyms come to be differentiated in meaning, one of them is usually lost".

[2]Original: "les synonymes n'existent pas longtemps : ou bien ils se différencient, ou bien l'un des deux termes disparaît".

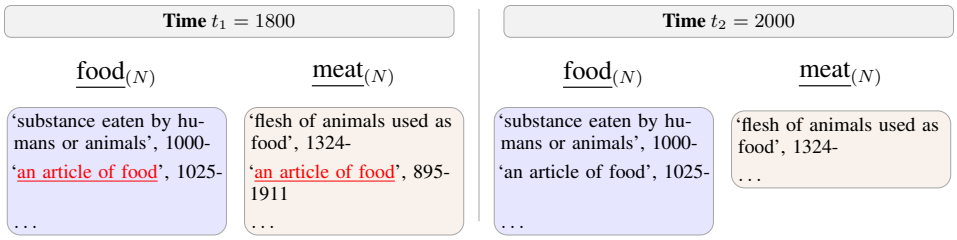


Figure 1: The lemmas food$_{(N)}$ and meat$_{(N)}$ each display several meanings at both $t_1 = 1800$ and $t_2 = 2000$. At $t_1$, both lemmas express the meaning 'an article of food', making the lemmas synonymous. At $t_2$, the lemmas stop sharing meanings (as meat$_{(N)}$ loses the shared meaning in 1911), and thus the two lemmas cease to be synonymous.

| Law | Category | Description |
|---|---|---|
| LD | **Synonyms Diverge** | *The lemmas cease to be synonymous.* |
| | DIFFERENTIATION | |
| | `Lost` | One or both lemmas do not exist at $t_2$. |
| | `Split` | No meanings are shared in the two lemmas at $t_2$. |
| LPC | **Synonyms Persist** | *The lemmas remain synonymous.* |
| | NO CHANGE | The lemmas contain the same meanings at both times. |
| | PARALLEL CHANGE | |
| | `Parallel Addition` | Each lemma gains $\geq$ 1 shared meaning(s) at $t_2$. |
| | `Parallel Loss` | Each lemma loses $\geq$ 1 shared meaning(s) at $t_2$. |
| | `Parallel Addition + Loss` | Each lemma both gains and loses $\geq$ 1 shared meaning(s) at $t_2$. |
| | NON-PARALLEL CHANGE | |
| | `Addition` | $\geq$ 1 lemma gains meaning(s) at $t_2$. |
| | `Loss` | $\geq$ 1 lemma loses meaning(s) at $t_2$. |
| | `Addition + Loss` | Both lemmas gain and/or lose meanings at $t_2$. |

Table 1: Taxonomy of evolution for synonymous lemmas between $t_1$ and $t_2$. In Synonyms Diverge, synonymous lemmas at $t_1$ cease to be synonymous at $t_2$ (DIFFERENTIATION), contributing to LD. In Synonyms Persist, the lemmas remain synonymous at $t_2$. Out of NO CHANGE, PARALLEL CHANGE, and NON-PARALLEL CHANGE, only PARALLEL CHANGE contributes to LPC. Examples from OED in Appendix A.

The LPC is attributed to Stern (1931, p. 227), who found a list of English adverbs whose meaning *shifted* from 'rapidly' to 'immediately', and stated: "It is natural that synonyms should be used in similar contexts, which are the expressions of similar mental content, and denote similar referents. It is then to be expected that the semantic consequences should be similar".[3]

**Revised Formulation** LD and LPC can be reformulated using the lemma as the central unit that can gain and/or lose meanings. This is exemplified by Figure 1, where food$_{(N)}$ and meat$_{(N)}$ cease to be synonymous at a later time, as the shared meaning 'an article of food' was lost from the latter lemma.

We revise LD and LPC through a fine-grained taxonomy of nine mutually exclusive leaf categories presented in Table 1. Given a pair of synonymous lemmas at $t_1$, at a later $t_2$, two broad outcomes can be distinguished.

Synonyms Diverge captures cases where lemmas that were synonymous at $t_1$ cease to be synonymous at $t_2$. This can happen either through the scenario where one or both of the lemmas are `Lost` in the later time period, or through `Split`, where one of the lemmas loses the meaning(s) that made the two lemmas synonymous. These categories contribute to the LD.

Synonyms Persist captures cases where the pair remains synonymous, with or without change. It consists of NO CHANGE, PARALLEL CHANGE, and NON-PARALLEL CHANGE, defined below. NO CHANGE covers cases in which a pair of synonymous lemmas has no change at all in their set of meanings between the two times. As the time difference $t_2 - t_1$ grows, we expect this class to consist of fewer and fewer pairs. This category does not contribute to either law. PARALLEL CHANGE covers cases where the same changes occur to both lemmas, either the gain (`Parallel Addition`), loss (`Parallel Loss`), or both gain and loss (`Parallel Addition + Loss`) of shared meaning. These categories contribute to the LPC. NON-PARALLEL CHANGE covers changes to the composition of either lemma that do not concern the shared meaning(s). That is, one or both of the synonymous lemmas can gain/lose one or more meanings *unrelated* to the shared meaning(s). These categories do not contribute to either law.

Each synonymous lemma pair is assigned to one, and only one, category in the taxonomy. For example, categories involving PARALLEL CHANGE take precedence over NON-PARALLEL CHANGE affecting other meanings in the lemma. For instance, if a pair undergoes `Parallel Addition` while also losing other unrelated meanings, it is classified only as `Parallel Addition`. This design ensures

[3] Note, however, that he also said: "Until more evidence is brought forward I think I may leave the question of synonym groups aside, as not proven".

| POS | Lemmas | Definitions | $\frac{\text{Def.}}{\text{Lemma}}$ ratio |
|---|---|---|---|
| ADJ | 45 471 | 83 182 | 1.83 |
| NOUN | 100 589 | 210 818 | 2.10 |
| VERB | 21 420 | 82 120 | 3.83 |
| Total | 167 480 | 376 120 | 2.25 |

Table 2: Dataset statistics per POS.

that the taxonomy captures the primary relationship governing the evolution of the synonymous pair.

## 4 Dataset Creation

To be able to test the prevalence of the two synonymy laws, we need a dataset that can evidence the prevalence of each. Previous work (§2) has taken raw corpora as a starting point, as it can provide empirical evidence on the usage of synonym pairs. While corpora can reveal the usages of senses in practice, we are left with the derivation of meaning through corpus samples, either via explicit word sense induction or by using embeddings. Both methods suffer from being imprecise, even more so as we wish to have time spans that go back in time hundreds of years (cf. §5).

We therefore employ the Oxford English Dictionary (OED, 1989), a high-quality resource in which meanings are explicitly defined and accompanied by their temporal span. While this does not provide us with usage statistics, we rely on the OED being a descriptive dictionary aimed at documenting actual language use.

We started from the full set of meaning definitions in the OED, totaling approximately 837 000 entries. We focused on adjectives (ADJ), common nouns (NOUN), and verbs (VERB). We collected definitions and example sentences using the OED Research API.[4] We then filtered out lemmas whose form was shorter than two characters, as well as definitions shorter than six words or shared by four or more lemmas, as these definitions tend to be overly general, uninformative, and formulaic (e.g., 'attributive, usually in figurative sense.'). See Table 2 for the dataset statistics after filtering.

### 4.1 Finding Synonymous Definitions

The OED provides definitions for the meanings of each lemma, but it does not establish mappings between those definitions. For example:

- 'Consisting of or characterized by dots; marked or covered with dots.' (dotty$_{(A)}$)
- 'Characterized by or marked with minute points or spots.' (puncticular$_{(A)}$)

Dotty$_{(A)}$ and puncticular$_{(A)}$ share a meaning and are therefore synonyms, but this meaning is expressed through two different definitions. Unlike a resource like WordNet, where synsets group synonyms, the OED neither has explicit linking nor offers any guarantee to use the same definition when two lemmas share a meaning.

Therefore, we must first identify pairs of definitions that convey the same meaning. This corresponds to the semantic textual similarity task (Guder et al., 2026) applied to definitions. Among the most effective approaches for this task is LLM-as-a-Judge, which mitigates one major limitation of embedding-based approaches that rely too heavily on surface-level lexical overlap (Nikolaev and Padó, 2023). However, relying on LLM-as-a-Judge to process the enormous number of instances (29 billion definition pairs) would be computationally infeasible. Thus, our pipeline is divided into two stages aimed at ensuring a trade-off between coverage, accuracy, and computational effort.

In the first stage (**Definition Embedding**), we aim to reduce the number of possible pairs by relying on definition embeddings, retaining only those whose cosine similarity is sufficiently high, that is, filtering out pairs that clearly do not express the same meaning. In the second stage (**Candidate Selection**), we refine the resulting set of definition pairs and keep only those that truly convey a similar meaning by using LLM-as-a-Judge in a few-shot setting. Finally, we perform a **Human Validation** on a subset of the resulting pairs (§4.2).

**Definition Embedding** We embedded all definitions using Qwen3-Embedding-8B (Zhang et al., 2025) using the prompt "Embed the following definition: `[DEFINITION]`". We selected this model because it ranked first and second (in April 2026), respectively, in the two MTEB benchmark (Muennighoff et al., 2023; Enevoldsen et al., 2024) tasks most relevant to our work: pair classification and semantic textual similarity.

For each POS, we computed cosine similarity between all pairs of embedded definitions. Table 8 in Appendix B provides a breakdown of these pairs by POS and cosine similarity bin. We note that the distribution for noun pairs peaks in the 0.35–0.40 similarity bin, while the peak for verb pairs is around 0.5, likely due to many meanings that increase the chances of overlap (cf. Table 2).

[4] Access was obtained through an academic license.

As expected, a high cosine similarity score is a suboptimal proxy of synonymy: definitions can be lexically very similar, while encoding substantially different meanings. For example, the definitions 'A person or thing which has been resurrected' (resurrection$_{(N)}$) and 'A person or thing which revivifies something or someone' (revivifier$_{(N)}$) received a cosine similarity score of 0.869, although they differ because of the agent vs. patient distinction.

**Candidate Selection** After removing pairs with low cosine similarity, we want to prevent false positives from being included in our dataset. Thus, we introduced an LLM-as-a-Judge to identify these pairs for removal from the set of definitions considered synonymous using a few-shot binary classification task. For each pair of definitions, and example sentences (if available), we asked the model to determine whether the two definitions were *similar* or *dissimilar*.[5] We used Qwen3.5-27B (Qwen Team, 2026) with the temperature set to 0 to ensure reproducibility.[6] See Figures 9–11 in Appendix D for the full prompt.

We randomly sampled 500 definition pairs per POS and cosine-similarity bin and evaluated the proportion deemed *similar* (i.e., synonymous) by Qwen3.5 at each cosine-similarity bin. Figure 2 displays, for each cosine-similarity bin, the total number of pairs in the full dataset alongside the synonymy rate estimated by Qwen3.5 derived in our sample. As expected, synonymy as estimated by Qwen3.5 is practically nonexistent below cosine similarity 0.50 (hence the $x$-axis range) and rises steadily thereafter; above 0.85, roughly 90% of pairs are synonyms according to Qwen3.5.

We would prefer a lower threshold so as not to miss potential synonym candidates, but that would dramatically increase the number of pairs that must be run through the LLM. Conversely, a higher threshold improves computational feasibility but risks discarding true synonyms. We therefore selected **0.80** as a pragmatic trade-off, yielding 169 573 ADJ, 419 845 NOUN, and 950 525 VERB definition pairs. Lowering the cutoff from 0.80 to 0.75, for example, would add roughly 3.85 million additional pairs to the LLM evaluation step, even though only about 59% of those extra pairs are estimated to be true synonyms, making the overall pipeline prohibitively expensive.

[5]We did not ask for *same/different* because we want to include pairs where the definitions refer to the same underlying meaning but differ in formality or register.

[6]Since licensing prevents us from using the OED data outside our own servers, we selected a model that could be reliably run on our GPUs.

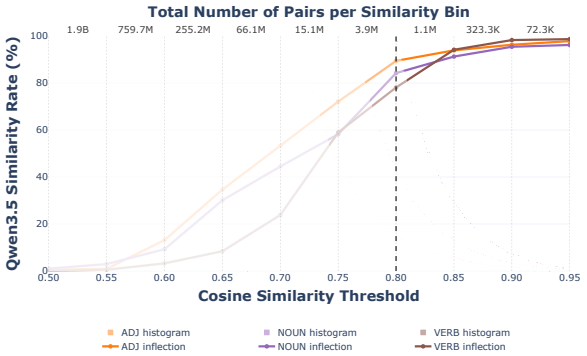


Figure 2: Trade-off between number of definition pairs and Qwen3.5 estimates of synonymous definitions per cosine-similarity bin.

We ran all definition pairs with a cosine similarity above 0.80 through Qwen3.5 and retained only those labeled as *similar*: 152 536 ADJ, 354 461 NOUN, and 798 631 VERB.

## 4.2 Human Validation

To validate the output of the **Candidate Selection** stage, we manually evaluated a subset of 300 definition pairs (100 per POS). We used three expert annotators (co-authors of this paper, experienced in linguistics). We adopted the same binary setup from §4.1: annotators were presented with forms, definitions, and example sentences, and were asked to judge whether the two definitions were *similar* (i.e., if they represent the same meaning), or *dissimilar* otherwise. For each POS, the sample comprised: 30 pairs with cosine similarity above 0.85 labeled similar by Qwen3.5; 30 pairs with cosine similarity above 0.85 labeled dissimilar by Qwen3.5; and 40 pairs with cosine similarity below 0.85 (also evaluated with Qwen3.5), drawn approximately uniformly from cosine similarities starting at 0.2.[7] The procedure allows us to assess both precision and recall of Qwen3.5 in classifying synonymous definitions. Attention checks were included to ensure data quality, and the order of items was randomized to avoid order effects. See Figure 9 in Appendix D for the instructions given to the annotators. The approximate completion time for the task was one hour per POS.

[7]The validation sample was drawn before the final operational threshold of 0.80 was determined (§4.1). We therefore used a stricter cosine-similarity boundary of 0.85 to construct evaluation strata with clear positive and negative examples.

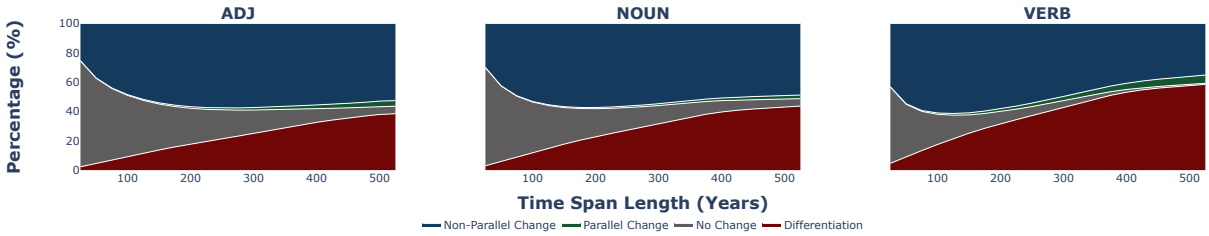


Figure 3: Aggregated taxonomic trends using pairs of synonymous lemmas, by POS and time-bin length. See Figure 5 in Appendix E for the full breakdown by leaf taxonomic category.

| POS | IAA ($\alpha$) | Acc. | Similar | | | Dissimilar | | |
|---|---|---|---|---|---|---|---|---|
| | | | Prec. | Rec. | F1 | Prec. | Rec. | F1 |
| ADJ | 0.73 | 0.93 | 0.84 | 0.97 | 0.90 | 0.98 | 0.91 | 0.94 |
| NOUN | 0.74 | 0.89 | 0.84 | 0.89 | 0.86 | 0.93 | 0.89 | 0.91 |
| VERB | 0.59 | 0.86 | 0.72 | 0.87 | 0.79 | 0.93 | 0.85 | 0.89 |

Table 3: Expert validation results per POS. | Classification metrics for Qwen3.5 predictions on the same data.

To assess consistency among the annotators, we computed inter-annotator agreement (IAA) using Krippendorff's $\alpha$ (Krippendorff, 1980). To assess Qwen3.5 performance, we computed precision, recall, F1, and accuracy by comparing the labels predicted by the model against gold labels defined as the majority vote of the three expert annotators for each pair.[8] Table 3 presents the results by POS. IAA results indicate moderate to substantial agreement among annotators, with VERB (0.59) lagging behind the other POS (0.73-0.74). The model's accuracy is consistently high across all POS, and the model achieves stronger F1 scores on the dissimilar pairs across all POS.

We repeated our study with non-expert annotators to compare their performance with that of our experts. 10 crowdworkers for each POS resulted in IAA between 0.41–0.58 and accuracy scores of 0.76–0.79, indicating that this is a hard task in need of expertise. See Appendix C for full details on the non-expert validation. Overall, the validation shows that the results from Qwen3.5 are reliable and accurate enough to be used for this study.

## 5 Analyses

**Experimental Setup** We considered the definitions in Table 2 together with their temporal span. Using this information, we could reconstruct the set of meanings of each lemma at any given time points $t_1$ and $t_2$, and then compare all lemmas to retain only those that are synonymous with at least one other lemma at $t_1$. Synonymy was determined using the methodology outlined in §4.1. For each pair of synonymous lemmas at $t_1$, we compared their sets of meanings at times $t_1$ and $t_2$, and assigned each pair to the appropriate category in the taxonomy presented in Table 1.

We applied this procedure in multiple time bins constructed by sampling start years $t_1$ at 25-year intervals from 1500 to 2000. We paired each $t_1$ with multiple end years $t_2$ at 25-year intervals, beginning at $t_1 + 25$ and extending up to 2025. This produced 231 valid $t_1 - t_2$ intervals with durations ranging from 25 to 525 years. Shorter intervals produced more bins because they could be generated from a larger number of valid starting years (e.g., 21 bins of 25-year spans, 20 bins of 50-year spans, etc.; only 1 bin of 525-year span). Counts in each taxonomic category were averaged across all bins for each interval length. All analyses are run for each POS individually.

**Main Results** Aggregated results at the intermediate taxonomic level (i.e., DIFFERENTIATION, NO CHANGE, PARALLEL CHANGE, and NON-PARALLEL CHANGE) can be found in Figure 3. We find that DIFFERENTIATION is far more prevalent than PARALLEL CHANGE, which is barely visible for any of the POS.

NO CHANGE, in which all meanings of two synonymous lemma pairs remain unchanged between $t_1$ and $t_2$, decreases over time, starting from a fairly high portion (between 52% for VERB and 72% for ADJ). That is, the longer the time span, the less likely that two synonymous lemmas remain completely unchanged. VERB is more likely to change overall, and in the 525-time span, only 0.74% of the pairs remained unchanged, compared to 5.1% for both ADJ and NOUN.

As for NON-PARALLEL CHANGE, it is rather frequent even at short time frames, with an incidence between 15% and 40% at 25 years. In the longer time bins, about 40-50% of the synonym

[8] We excluded pairs for which annotators did not reach a majority vote: 3 ADJ, 8 NOUN, and 3 VERB.

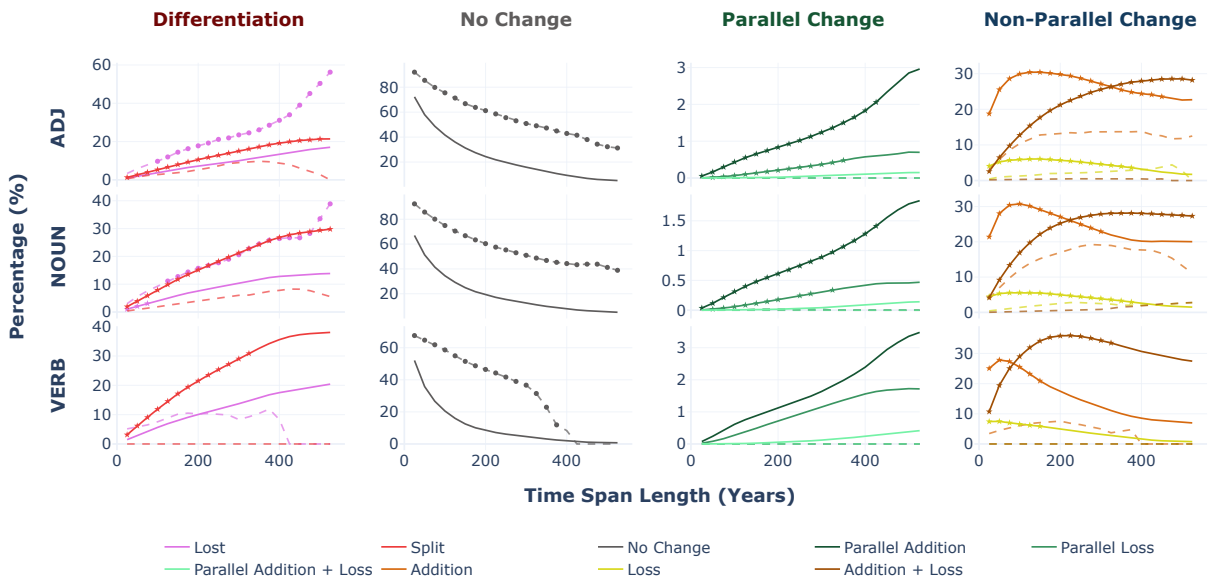


Figure 4: Synonym (solid lines) vs. control (dashed lines) sets, broken down by POS and time-bin length. Y-axes use different scales due to differences in percentage ranges across categories. Stars and dots are shown for each 25-year span and indicate statistically significant differences between the synonym and control setups ($p < 0.05$). A star indicates that the effect of the law is stronger in the synonym setup, while a dot that the effect is stronger in the control setup. The absence of a marker indicates no statistically significant difference for that time bin.

pairs add or lose some meanings that do not affect the synonymy relation.

**Control Condition** To investigate whether our results are statistically significant, we introduced a control condition. We sampled 500 000 definition pairs with cosine similarity below 0.80 and filtered them to remove any lemma that ever participates in a synonymy relation at any point in time. From each remaining lemma pair, we randomly selected one meaning per lemma[9] to form an artificial mock-synonym pair, and processed it through the same pipeline as true synonyms. We note that a large proportion of lemmas participate in at least one synonymy relation (cf. Table 4).

Because non-synonymous lemma pairs lack shared meaning(s), for the `Split` and `Parallel Loss` categories we tracked the eventual disappearance of the matched meaning at time $t_2$. The control group should exhibit no `Parallel Addition` or `Parallel Addition + Loss`; it therefore provides a baseline against which we test whether the amount of these categories in synonyms is significantly higher than 0.

For each time span (i.e., 25 years, 50 years, . . . , 525 years), we computed a chi-square test[10] to

[9]We repeated this random selection twice more to verify consistency; both runs produced comparable results.

[10]We used Fisher's exact test whenever expected frequencies were below 5.

| POS | Set | Total | Monosemous | Polysemous |
|---|---|---|---|---|
| ADJ | Synonym | 31 126 | 19 322 | 11 804 |
| | Control | 13 051 | 11 663 | 1 388 |
| NOUN | Synonym | 65 842 | 39 052 | 26 790 |
| | Control | 21 480 | 18 403 | 3 077 |
| VERB | Synonym | 19 869 | 9 305 | 10 564 |
| | Control | 356 | 317 | 39 |

Table 4: Distribution of lemmas across synonym and control sets with polysemy information per POS.

assess whether the distributions of the laws differed significantly between the synonym and control groups. All $p$-values were corrected for multiple comparisons using the Benjamini–Hochberg procedure. To determine the direction of the effect, i.e., whether a law was more prevalent in the synonym or control group, we computed odds ratio.

Figure 4 summarizes the main findings of the comparison to the control. Appendix E provides additional comparisons.

First, we observe that the distribution of categories is broadly comparable across POS (compare across rows). When comparing the synonym and control sets, we find that NO CHANGE is significantly more prevalent in the control group, suggesting that non-synonymous lemma pairs are much more stable than synonyms. This pattern is further supported by the higher prevalence of NON-PARALLEL CHANGE in the synonym group, indicating that synonyms are generally more dynamic.

| | ADJ | | NOUN | | VERB | |
|---|---|---|---|---|---|---|
| **Category** | **Poly.** | **Morph.** | **Poly.** | **Morph.** | **Poly.** | **Morph.** |
| Lost | 2.65 | 4.11 | 2.52 | 2.65 | 5.31 | 0.25 |
| Split | 5.96 | 1.39 | 5.92 | 1.13 | 10.60 | 0.02 |
| Parallel Addition | 5.52 | 15.01 | 5.41 | 18.72 | 14.35 | 0.31 |
| Parallel Loss | 8.05 | 13.18 | 7.59 | 11.23 | 16.94 | 0.22 |
| Parallel Addition + Loss | 5.30 | 31.16 | 4.71 | 24.18 | 17.79 | 0.44 |

Table 5: Poly.: mean number of definitions per lemma pair at $t_1$ averaged across all synonym pairs. Morph.: percentage of lemma pairs sharing the same root.

For DIFFERENTIATION, we see that Split is more prevalent in the synonym group, indicating that synonymous lemmas tend to diverge over time, whereas Lost is more prevalent in the control group. This suggests that participating in a synonymy relation protects a lemma from being lost. For PARALLEL CHANGE, while we have no formal control group, we find that Parallel Addition and Parallel Loss are (for about half of the intervals) more than 0, while Parallel Addition + Loss is never significantly higher than 0.

**Other Factors** During our analyses, we observed potential effects related to both polysemy and morphology. For example, many lemma pairs classified as Lost involved nearly monosemous lemmas, whereas pairs which share the same lexical root, frequently underwent Parallel Addition. We therefore systematically investigated the extent to which polysemy and morphology influence LD and LPC. For polysemy, for each synonymous lemma pair, we computed the average number of meanings across the pair at $t_1$. For morphology, we compared the roots of the two lemmas in each synonymous pair. We first segmented lemmas into morphemes using the NLTK Snowball Stemmer (Bird and Loper, 2004). If the roots were identical (e.g., pestilenti-al and pestilenti-ous), the pair was considered morphologically related. The results are shown in Table 5.

We notice two strong effects. Firstly, the polysemy values of the Lost category are significantly lower than those of other categories. If a lemma has very few meanings at $t_1$, it is more likely to disappear completely by $t_2$ if it loses a meaning than a highly polysemous lemma, leading to Lost. Secondly, we note that the categories participating in the LPC have a significantly higher degree of morphological overlap, especially for ADJ and NOUN. The percentages for VERB are low because verb pairs rarely share the same root.

## 6 Discussion and Conclusion

We examined two contrasting laws in historical linguistics concerning the evolution of synonyms: the LD and LPC. We developed a lemma-based framework, decomposing the laws into a fine-grained taxonomy that allows us to address critical nuances overlooked in prior work.

We interpret the LPC as referring specifically to PARALLEL CHANGE, using our taxonomy to exclude NO CHANGE and NON-PARALLEL CHANGE. The results in Figure 3 motivate this interpretive distinction. In fact, the majority of the class Synonyms Persist correspond to NO CHANGE and NON-PARALLEL CHANGE. What remains of PARALLEL CHANGE is significantly non-zero in about half of the bins (Figure 4) for ADJ and NOUN and never for VERB. A significant fraction of the observed PARALLEL CHANGE is driven by lemmas sharing a root (Table 5).

The prevalence of DIFFERENTIATION is higher than PARALLEL CHANGE and increases the longer the time interval. In absolute terms, DIFFERENTIATION covers about 44% of the synonym pairs for NOUN, 38% for ADJ, and 58% for VERB at the longest interval of 525 years (Figure 3). However, when we account for the control group, we see that Lost is significantly higher in the control (Figure 4), and that being a synonym reduces the likelihood of one or both of the lemmas to disappear. One could argue that LD should include only Split. However, we also note that lemmas that have at least one synonymy relation have significantly more meanings on average than the control lemmas (Table 4). Thus, the fact that Lost exceeds Split in the control while the reverse holds in the synonym group is likely attributable to the degree of polysemy: if a highly polysemous lemma loses a meaning, then the pair to which it belongs will Split; if a monosemous lemma loses a meaning, then the lemma will be Lost. When considering both categories of LD across the synonym and control sets (Figure 4), we find that there is no statistically significant difference in the mean. Thus, DIFFERENTIATION happens as often in the control group as it would on random, and this suggests that the LD is not driven by a property of synonyms but by the general tendency for meanings to get lost.

In summary, our results support LD over LPC, but maintain that LD is driven by the general tendency for meanings to get lost, and that LPC is influenced by synonyms that share a root.

## Limitations

We applied our framework to English only. Applying it to other languages is an interesting direction for future work, as it would allow us to investigate whether the LD or the LPC is predominant across languages. Applying it to other resources allows us to detect if the behavior we are observing is driven by lexicographic choices, rather than a *true* language feature. The main challenge in both cases lies in identifying resources that are comparable to the OED.

Previous work relied exclusively on unsupervised methodology for extracting meaning from corpora. The methodology has some disadvantages, in particular if one relies on type vectors. Our work instead relies on a high-quality, expert resource carefully defining word meaning. Nonetheless, these are complementary views, and a full study should combine the two and ground lexical resources in empirical evidence from corpora. Because, while the OED might list lemmas as synonymous because they share a meaning, it could very well be that in practice, the lemmas are rarely or never used synonymously.

We did not have access to a historical thesaurus (e.g., the Historical Thesaurus of the OED), which would have offered an independent, directly encoded representation of synonymous relations. Note, however, that large thesauri typically represent broad ontological domains only partially relevant to the specific phenomenon of synonymy (Lietard et al., 2023). In addition, modeling synonymy through definitions provides a relevant and challenging test case for current computational models, and allows for richer and more flexible analyses than what could be extracted from the traditional structure of a thesaurus.

Because adjacent temporal windows overlap substantially, observations across bins are not independent. Consequently, statistical tests over time-bin aggregates should be interpreted cautiously; the overarching diachronic patterns nonetheless remain consistent.

## Ethical Considerations

The 30 crowdworkers recruited via Prolific were compensated above the platform's minimum pay requirements, for a total of £252. All participants provided informed consent prior to taking part in the experiment.

All experiments were conducted on compute nodes equipped with 4× NVIDIA A100 GPUs, for a total of 4,525 hours of computation.

## References


Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Michel Bréal. 1897. *Essai de Sémantique*. Hachette, Paris.

Yingying Cai and Hendrik De Smet. 2024. Are categories' cores more isomorphic than their peripheries? *Frontiers in Communication*, Volume 9 - 2024.

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.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, and 63 others. 2024. MMTEB: Massive Multilingual Text Embedding Benchmark. In *The Thirteenth International Conference on Learning Representations*.

Mariia Fedorova, Timothee Mickus, Niko Partanen, Janine Siewert, Elena Spaziani, and Andrey Kutuzov. 2024. AXOLOTL'24 Shared Task on Multilingual Explainable Semantic Change Modeling. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 72–91, Bangkok, Thailand. Association for Computational Linguistics.

Karlien Franco, Mariana Montes, and Kris Heylen. 2022. Deconstructing destruction: A cognitive linguistics perspective on a computational analysis of diachronic change. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 23–32, Dublin, Ireland. Association for Computational Linguistics.

Matt Hunt Gardner, Eva Uffing, Nicholas Van Vaeck, and Benedikt Szmrecsanyi. 2021. Variation isn't that hard: Morphosyntactic choice does not predict production difficulty. *PLOS ONE*, 16(6):1–39.

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.

Larissa Guder, João Paulo Aires, Hígor Uélinton da Silva, Felipe Meneguzzi, and Dalvan Griebler. 2026. Sentence representations for semantic textual similarity: A systematic review. *Computer Speech & Language*, 100:101970.

John Haiman. 1980. The iconicity of grammar: Isomorphism and motivation. *Language*, pages 515–540.

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

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA. Section: 12.

Benoît Leclercq and Cameron Morin. 2023. No equivalence: A new principle of no synonymy. *Constructions*, 15(1).

Natalia Levshina and David Lorenz. 2022. Communicative efficiency and the principle of no synonymy: predictability effects and the variation of want to and wanna. *Language and Cognition*, 14(2):249–274.

Bastien Lietard, Mikaela Keller, and Pascal Denis. 2023. A Tale of Two Laws of Semantic Change: Predicting Synonym Changes with Distributional Semantic Models. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 338–352, Toronto, Canada. Association for Computational Linguistics.

Saif M. Mohammad. 2020. WordWars: A Dataset to Examine the Natural Selection of Words. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3087–3095, Marseille, France. European Language Resources Association.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Dmitry Nikolaev and Sebastian Padó. 2023. Representation biases in sentence transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3701–3716, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Francesco Periti and Stefano Montanelli. 2024. Lexical Semantic Change through Large Language Models: a Survey. *ACM Computing Surveys*.

Francesco Periti and Nina Tahmasebi. 2024. Towards a complete solution to lexical semantic change: an extension to multiple time periods and diachronic word sense induction. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 108–119, Bangkok, Thailand. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

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.

Gustaf Stern. 1921. *Swift, swiftly, and their synonyms: A contribution to semantic analysis and theory*. Wettergren & Kerber.

Gustaf Stern. 1931. *Meaning and Change of Meaning: With Special Reference to the English Language*. Indiana University Press.

Edgar Howard Sturtevant. 1917. *Linguistic change: An introduction to the historical study of language*. University of Chicago Press.

Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, Simon Hengchen, Ekaterina Vylomova, Nick Haslam, Sampriti Mahanty, Frank Boons, Julia Handl, Riza Batista-Navarro, Stellan Petersson, Emma Sköldberg, Zheng-sheng Zhang, Karan Grewal, Ana-Sabina Uban, Alina Maria Ciobanu, Liviu P. Dinu, Yi-jun Duan, and 8 others. 2021. *Computational approaches to semantic change*. Language Science Press.

Peter D. Turney and Saif M. Mohammad. 2019. The natural selection of words: Finding the features of fitness. *PLOS ONE*, 14(1):e0211512.

Ana-Sabina Uban, Alina Maria Ciobanu, and Liviu P. Dinu. 2021. Cross-lingual laws of semantic change. *Computational approaches to semantic change*, pages 219–260. ISBN: 9783961103126. Berlin. Language Science Press.

Yang Xu and Charles Kemp. 2015. A Computational Evaluation of Two Laws of Semantic Change. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, volume 37, pages 2703–2708, Pasadena, California. Cognitive Science Society.

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.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint*. ArXiv:2506.05176 [cs].

## A Details on Framework Categorization

Table 7 presents example lemma pairs and the explanations for the categorization of each pair based on the taxonomy in Table 1. The examples are a subset retrieved from applying our methodology (§5) at times $t_1 = 1800$ and $t_2 = 2000$.

## B Additional Dataset Details

Table 8 shows the total number of definition pairs in our dataset, per cosine similarity bin and per POS.

## C Validation by Non-Experts

We presented the same task used with expert annotators in §4.2 to non-expert participants recruited via Prolific, in order to compare the performance of non-linguists with that of experts. We recruited 30 crowdworkers, allocating 10 participants per POS. All participants were native English speakers, born in the UK. The average completion time was approximately 40 minutes.

Table 6 presents the results. Scores are comparable across POS. In terms of consistency, IAA is moderate. Using the gold labels from non-experts, we observe that Qwen3.5 performance decreases both in overall accuracy and across individual classes with respect to that of experts (cf. Table 3). Overall, the consistently lower IAA relative to expert annotations suggests that this is a challenging task that is not reliably feasible for lay annotators.

| POS | IAA ($\alpha$) \| | Acc. | Similar | | | Dissimilar | | |
|---|---|---|---|---|---|---|---|---|
| | | | Prec. | Rec. | F1 | Prec. | Rec. | F1 |
| ADJ | 0.41 | 0.79 | 0.92 | 0.67 | 0.78 | 0.70 | 0.93 | 0.80 |
| NOUN | 0.58 | 0.76 | 0.97 | 0.61 | 0.75 | 0.62 | 0.97 | 0.76 |
| VERB | 0.49 | 0.77 | 0.94 | 0.63 | 0.76 | 0.67 | 0.95 | 0.79 |

Table 6: Crowdworker validation results per POS. | Classification metrics for Qwen3.5 predictions on the same data.

## D Instructions and Prompts

We provide instructions (Figure 9) and prompts (Figures 10 and 11) used in this study.

## E Analyses: Detailed Results

We report the detailed results of the analyses performed in §5. Figure 5 accompanies Figure 3, with a breakdown of the results, showing means and standard deviations by leaf taxonomic category. We also show detailed results, both aggregated and by leaf taxonomic categories, comparing synonym and control groups, per POS (Figure 6 for ADJ, Figure 7 for NOUN, and Figure 8 for VERB).

| Category | Example Pair | Explanation |
|---|---|---|
| Lost | hontous$_{(A)}$, shameful$_{(A)}$ | At $t_2$, hontous$_{(A)}$ does not exist anymore, while shameful$_{(A)}$ continues to exist. |
| Split | food$_{(N)}$, meat$_{(N)}$ | At $t_2$, meat$_{(N)}$ loses the meaning 'a kind of food; an article of food, a dish', leaving no meaning which is shared with food$_{(N)}$. |
| NO CHANGE | endless$_{(A)}$, eternal$_{(A)}$ | Between $t_1$ and $t_2$, both lemmas retain exactly the same set of meanings. |
| Parallel Addition | disguise$_{(V)}$, dissimulate$_{(V)}$ | At $t_2$, both lemmas gain new shared meaning 'to conceal the presence of (electricity) by neutralization'. |
| Parallel Loss | regression$_{(N)}$, retrogradation$_{(N)}$ | At $t_2$, both lemmas lose the shared meaning 'the action or an act of going or coming back'. |
| Parallel Addition + Loss | equalize$_{(V)}$, even$_{(V)}$ | At $t_2$, both lemmas gain new (specialized) shared meaning 'to make the score (in a sporting contest) equal to that of the opponent', and lose the (general) shared meaning 'to make or become equal'. |
| Addition | miserable$_{(A)}$, wretched$_{(A)}$ | At $t_2$, miserable$_{(A)}$ gains two new meanings, 'causing great discomfort' and 'of tea: very weak'. |
| Loss | disclose$_{(V)}$, reveal$_{(V)}$ | At $t_2$, disclose$_{(V)}$ loses four meanings, and reveal$_{(V)}$ loses one meaning. |
| Addition + Loss | organism$_{(N)}$, whole$_{(N)}$ | At $t_2$, organism$_{(N)}$ gains two meanings, and whole$_{(N)}$ loses one meaning. |

Table 7: Explanations for the categorization of each example lemma pair in the taxonomy of synonymy evolution illustrated in Table 1 at $t_1 = 1800$ and $t_2 = 2000$. Example pairs are synonymous at $t_1$.

| POS \ BIN | [0,.05) | [.05,.10) | [.10,.15) | [.15,.20) | [.20,.25) | [.25,.30) | [.30,.35) | [.35,.40) | [.40,.45) | [.45,.50) | [.50,.55) | [.55,.60) | [.60,.65) | [.65,.70) | [.70,.75) | [.75,.80) | [.80,.85) | [.85,.90) | [.90,.95) | [.95,1] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ADJ | 0 | 0 | 57 | 15.3K | 1.10M | 33.1M | 306M | 936M | **1.16B** | 702M | 239M | 58.7M | 13.7M | 3.53M | 1.04M | 344K | 119K | 37.9K | 10.0K | 2.5K |
| NOUN | 0 | 7 | 9.9K | 1.71M | 84.3M | 1.18B | 4.87B | **7.45B** | 5.40B | 2.30B | 697M | 178M | 44.6M | 11.8M | 3.34M | 998K | 306K | 88.1K | 21.0K | 4.2K |
| VERB | 0 | 0 | 0 | 10 | 5.3K | 497K | 13.5M | 133M | 541M | **982M** | 917M | 523M | 197M | 50.7M | 10.7M | 2.51M | 707K | 197K | 41.3K | 4.9K |

Table 8: Total number of definition pairs per cosine similarity bin per POS.

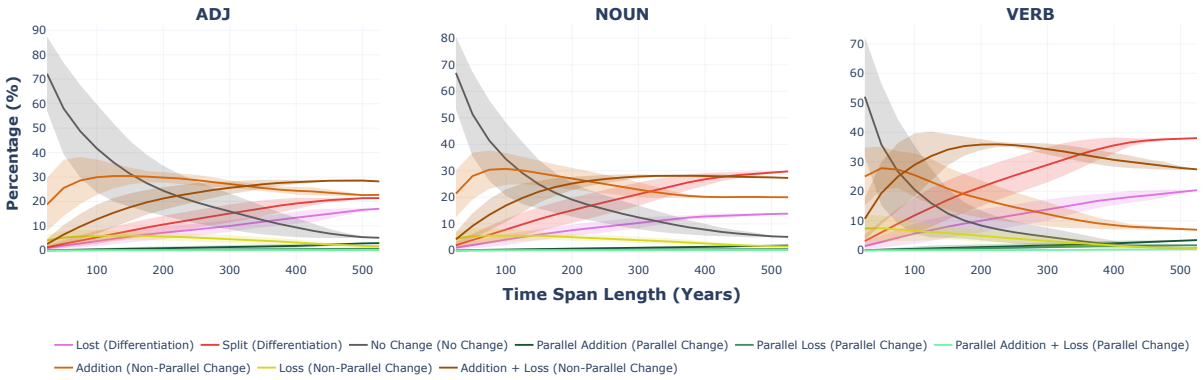


Figure 5: Full breakdown of mean and standard deviation of change percentages for the synonym group by leaf taxonomic category versus time span length, by POS, accompanying Figure 3.

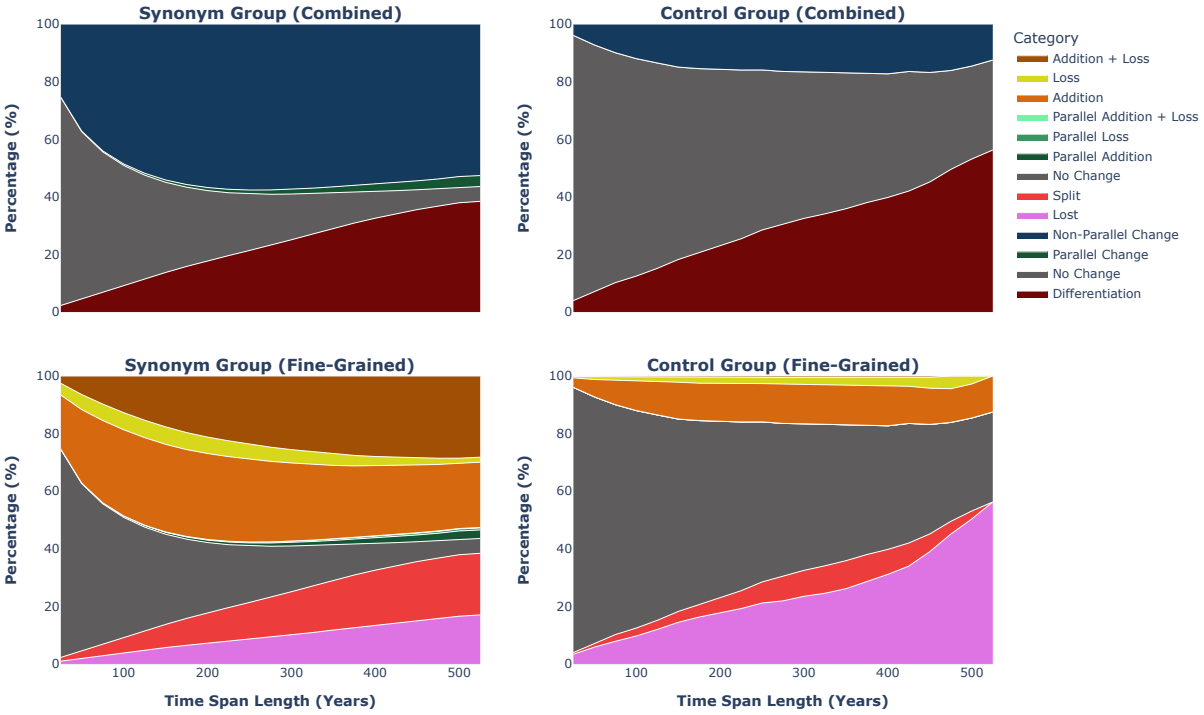


Figure 6: Detailed results of the analyses performed in §5, comparing synonym and control groups for ADJ. Top: aggregated categories. Bottom: leaf categories shown individually. We see clearly the significantly higher amount of stable lemma pairs in the control group (NO CHANGE). Conversely, we observe much higher NON-PARALLEL CHANGE in the synonymy group, particularly, `Addition + Loss`. We also observe that combining `Lost` and `Split` into DIFFERENTIATION shows similar sizes and behavior across synonym and control groups, but that the control group presents more `Lost`, while the synonym group presents more `Split`.

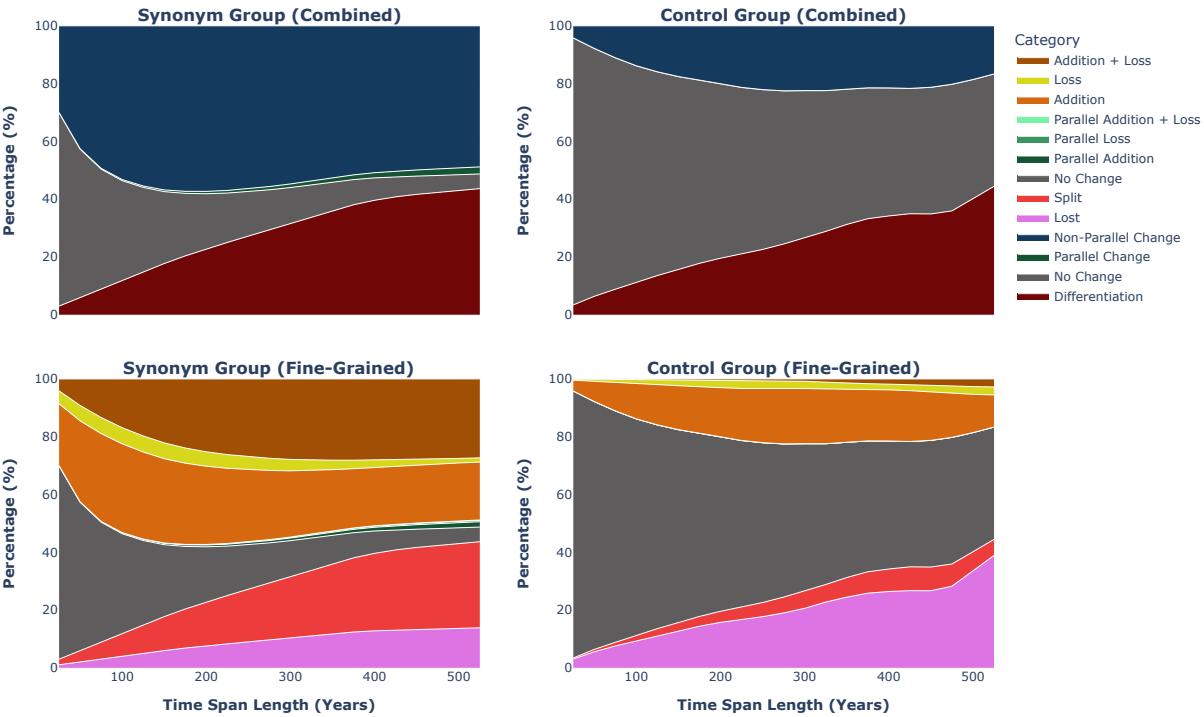


Figure 7: Detailed results of the analyses performed in §5, comparing synonym and control groups for NOUN. Top: aggregated categories. Bottom: leaf categories shown individually. NOUN behaves very similarly to ADJ, so we refer to Figure 6 for a detailed description.

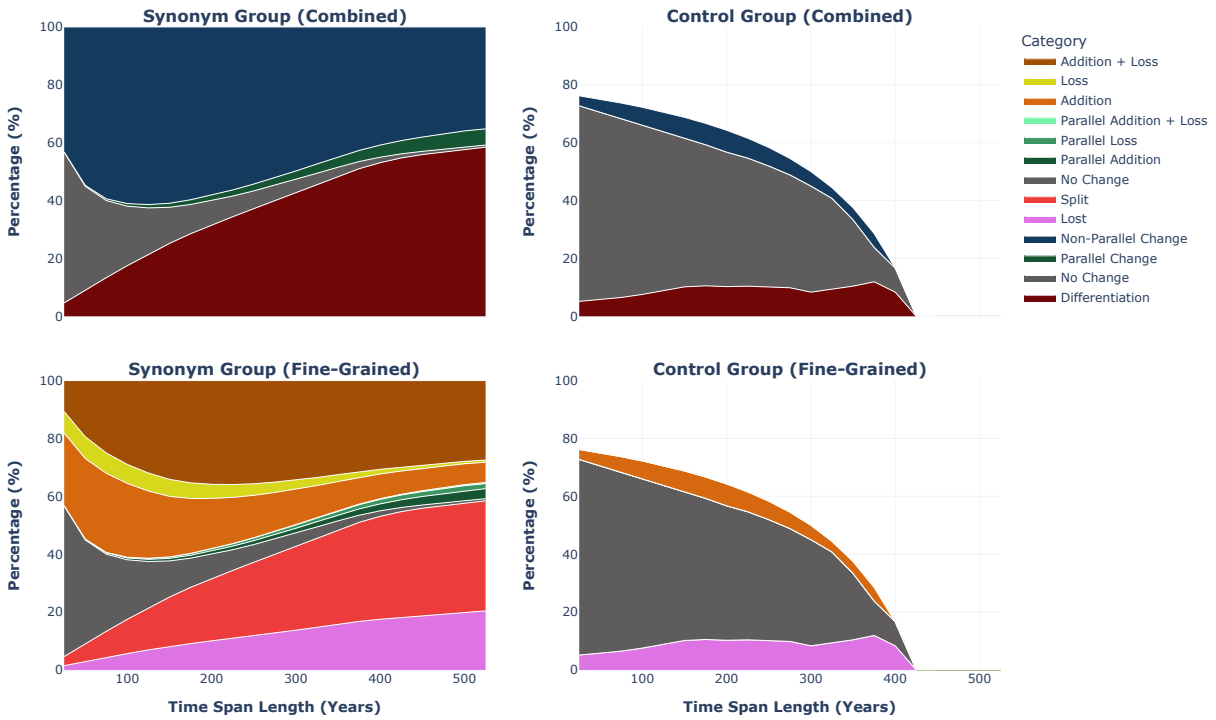


Figure 8: Detailed results of the analyses performed in §5, comparing synonym and control groups for VERB. Top: aggregated categories. Bottom: leaf categories shown individually. Verbs are special as 98% of all the verbs participate in a synonymy relation and thus the control group is very limited in size (cf. Table 4). Conclusions are therefore difficult to draw with respect to the control group. Compared to the other POS, we note a larger `Split` category among the synonyms, which we attribute to the higher number of meanings per lemma for verbs.

The task consists of rating the meaning similarity between two dictionary definitions. The meanings can be **SIMILAR** or **DISSIMILAR**. If the relationship between the definitions is unclear or ambiguous, you may select **CANNOT DECIDE**, but use this option sparingly.
When making your judgment, please consider the overall meaning of the definitions, rather than focusing on specific words or phrases.

**Examples:**

**SIMILAR** – The two definitions describe the same or closely related meanings.
**Example 1**
Word 1: dotty
Definition 1: "Consisting of or characterized by dots; marked or covered with dots; dotted."

Word 2: puncticular
Definition 2: "Characterized by or marked with minute points or spots."

**Example 2**
Word 1: fatty
Definition 1: "Full of fertilizing matter. Of soil: Fat, rich."

Word 2: gleby
Definition 2: "Of soil: Full of clods; rich, fertile."

**DISSIMILAR** – The two definitions have weak or no connection, or describe opposite concepts.
**Example 1**
Word 1: mutilating
Definition 1: "That mutilates a person or thing; causing mutilation or disfigurement."

Word 2: torturing
Definition 2: "That tortures; inflicting or causing torture; tormenting, excruciating."

**Example 2**
Word 1: monogynous
Definition 1: "(Entomology) (Of a colony of social insects) having only one queen."

Word 2: polygyne
Definition 2: "(Entomology) (Of a colony of social insects) having more than one queen."

When performing the task, you will also be shown example sentences to help you make your decision, in case the definitions are not sufficient.

Figure 9: Instructions given to both human annotators for the validation task and Qwen3.5 LLM-as-a-Judge.

You are an experienced lexicographer, expert in semantic relationships. Follow the instruction exactly.

Figure 10: System prompt used for Qwen3.5 LLM-as-a-Judge.

Now label the following pair.

Word 1: {word1}
Definition 1: {definition1}
Sample sentence 1: {sample_sentence1}

Word 2: {word2}
Definition 2: {definition2}
Sample sentence 2: {sample_sentence2}

Return exactly one label from: SIMILAR, DISSIMILAR, CANNOT DECIDE.
Output format: **LABEL: <one label>**
Do not include explanations, steps, or any other text.

Figure 11: Additional prompt postpended after the instructions in Figure 9 for Qwen3.5 LLM-as-a-Judge.