Using Synchronic Definitions and Semantic Relations to Classify Semantic Change Types

Pierluigi Cassotti🌷 , **Stefano De Pascale**🌿 , **and Nina Tahmasebi**🌷

🌷University of Gothenburg, 🌿VUB/FWO/KU Leuven

{pierluigi.cassotti,nina.tahmasebi}@gu.se

stefano.de.pascale@vub.be

Abstract

There is abundant evidence of the fact that the way words change their meaning can be classified in different *types* of change, highlighting the relationship between the old and new meanings (among which generalization, specialization and co-hyponymy transfer). In this paper, we present a way of detecting these types of change by constructing a model that leverages information both from synchronic lexical relations and definitions of word meanings. Specifically, we use synset definitions and hierarchy information from WordNet and test it on a digitized version of Blank's (1997) dataset of semantic change types. Finally, we show how the sense relationships can improve models for both approximation of human judgments of semantic relatedness as well as binary Lexical Semantic Change Detection.

1 Introduction

At any point in time, a word can have several meanings. Often, these meanings share a certain degree of relationship with one another and the word is then said to display *polysemy*. Diachronically, polysemy is a result of *semantic change*, the process in which a word change its meaning/s. When new meanings are derived, the relationship to existing meanings is determined by the type of change that takes place. Examples include metaphorical extensions (*the **arm** of the sea*) and specialization of usage (*sand* that used to mean the shore as well as the grain of sand, and now only the latter). Several taxonomies of semantic change have been proposed, from Reisig (1839) to Blank (1997), that describe both the type of change as well as the causes of it.

However, the computational community, which has spent the past 15 years developing computational models for automatically detecting semantic change in corpora, has disregarded those taxonomies and processes proposed in the past (Tahmasebi et al., 2021). Sparked by the invention of neural word embedding techniques, the focus has been on quantifying the *degree of change*, or determining *when* change has taken place, without the use of sense-inventories, dictionaries or knowledge-bases. Currently, the standard recipe for lexical semantic change detection (LSC) is form-based (Montanelli and Periti, 2023): each usage of a word is represented by its contextualized embedding, and the embeddings of different time periods are compared using cosine similarity. The comparison is done either by averaging pairwise similarities, or first averaging the embeddings of each time periods and then comparing the averages. When sense-based models are employed, (a) the embeddings are clustered to derive sense representations and then clusters are compared over time, and optionally (b) the clusters are labeled to ease interpretation.

So far, no computational work has been done to *detect the qualitative type of change* that words experience. Nor have these change types been utilised when building models for change detection. Our work therefore presents a first approach to addressing both of these tasks.

We rely directly on generated *definitions* and sidestep the standard recipe, i.e., to represent each usage of a word with its contextualized embedding. We consider a word's meaning, i.e., sense, as a *latent construct* and the definition as one way to capture that latent construct. The use of definitions to precisely describe the meaning of a word offers an ideal level of sense representation, that combines both precision and abstraction, as well as a way to summarize multiple usages. Thus, the use of definitions serves as a dimension reduction technique, projecting N×M contextualized embeddings in two time periods to a significantly smaller set of n×m definitions. By using a single computational step rather than two (steps (a) and (b) above), we reduce the complexity and decrease the uncertainty.

To classify semantic change types, we train a classifier on a *synchronic* sense repository with def-

initions and semantic relations, namely WordNet. The rationale behind this choice is that each of the type of semantic change targeted in this paper corresponds to a synchronic relationship between senses. This distant training procedure via synchronic information is necessary as there are no available resources to date for change type classification that surpass a few hundred examples.

Next, we test this classifier on an unseen dataset of historical semantic changes. We therefore extend the dataset provided by Blank (1997), in which for a given word and a pair of senses there is an associated type of semantic change, with concise definitions for each sense. Finally, we incorporate change type classification in a state-of-the art model for detection of semantic change, and show an improvement in the model performance.

Contributions: In this paper, we show that
1. definitions of word senses can be used to detect semantic change type;
2. we can classify the type of semantic change by training on synchronic sense relationships using sense definitions; and that
3. the use of type information can improve models for both graded Word-In-Context (WiC) as well as semantic change detection.

2 Background and Related Work

2.1 Lexical Semantic Change Detection

Lexical Semantic Change Detection (LSC) is the task of automatically identifying words that change their meaning over time. Meaning changes that take place because whole senses are added or lost over time are considered as *binary change*. Change can also be more subtle and relate to existing senses, for example, the decrease in frequency of a dominant sense. We consider the latter as *graded change*. Words can experience both binary and graded change over (longer or shorter) periods of time, and both add some senses, loose some, and change others.

Up until 2020, the majority of the methods that computationally modeled semantic change were tested only on the ability to say if a word had changed or not, conflating all kinds of change into one. Evaluation was ad-hoc and as a result, no comparability across different methods was possible. This changed when the SemEval-2020 Task 1 on Unsupervised Lexical Semantic Change Detection campaign (Schlechtweg et al., 2020) was launched in 2020. As a part of the shared task, benchmarks

covering Swedish, English, German and Latin were released. Each benchmark covers a set of words, across a pair of corpora (or subcorpora) with a temporal gap between the two. For each word, up to a 100 usages are sampled from each time period. These usages are paired both within a corpus and across the corpora. Each annotated pair receive a *graded* Word-in-Context score (see Table 1) by human annotators. The usages are considered *nodes* and the pairwise annotations as *edges* when a diachronic word usage graph (DWUG) (Schlechtweg et al., 2021) is created. The DWUG is clustered using correlation clustering and distribution of cluster members across time is used to attain the binary and graded change scores for each word. The information within a corpus facilitates the possibility to differentiate the synchronic senses (polysemy and homonymy), while comparing across corpora for the estimation of semantic change.

4:	Identical	Identity
3:	Closely related	Context variance
2:	Distantly related	Polysemy
1:	Unrelated	Homonymy

Table 1: The DURel relatedness scale (Schlechtweg et al., 2018) and the respective Continuum of semantic proximity proposed by Blank (1997).

Methods for encoding word meaning in the LSC task range from distributional semantic models such as count-based or Word2Vec approaches (Hamilton et al., 2016) and contextualized models (Laicher et al., 2021) to the most recent task-informed models such as XL-LEXEME (Cassotti et al., 2023) and models based on definitions (Giulianelli et al., 2023). Task-informed models are supervised models leveraging information from different tasks for LSC prediction. XL-LEXEME and Deep Mistake (Arefyev et al., 2021) are models trained on the WiC task, while Gloss Reader (Rachinskiy and Arefyev, 2021) exploits the Word Sense Disambiguation task. Definition-based models, on the other hand, finetune LLMs for in-context definition generation. XL-LEXEME holds the state of the art in multiple languages (Periti and Tahmasebi, 2024) and will thus be used here.

2.2 Sense Definition Modeling

Highly related to our work is the task of Word Sense Disambiguation (WSD) that has long been of great interest to the community (Navigli, 2009). Similar to other NLP tasks, model architectures

have evolved over time, including models based on knowledge-bases or the more common ones based purely on neural networks. With the outstanding performance on a majority of existing NLP tasks, generative Large Language Models (LLMs) have contributed to a paradigmatic shift also for WSD. While previous models assign a pre-defined label to a given usage of word w, current models generate the label of the word meaning. In particular, such a label could be the definition of the word in the specific usage context.

In this work we refer to *definition modeling* or *definition generation* as the task of automatically generating dictionary-like definitions of a (latent) word sense given a certain a word usage. The task is carried out using several different approaches (Gardner et al., 2022) and datasets, such as Word-Net (Fellbaum, 1998), Wiktionary [1], Oxford English Dictionary (OED) [2] and the Urban Dictionary [3]. Noteworthy is the contribution of Bevilacqua et al. (2020), which is one of the first works using a large pre-trained model, followed by Huang et al. (2021) who introduce a way to control the specificity of the generated sense definitions, and finally, Giulianelli et al. (2023) who use similarities between word sense definitions to approximate human judgments on semantic proximity. We will use the latter model in our work.

2.3 Classification in Historical Semantics

The way in which semantic change manifests itself is of great interest to the linguistic community and over time various taxonomies of semantic change have been proposed (Reisig, 1839; Paul, 1880; Darmesteter, 1893; Bréal, 1904; Stern, 1931; Bloomfield, 1933; Ullmann, 1957; Blank, 1997). These include a rich set of examples that have undergone semantic change. While there is a large overlap between the taxonomies (with generalization, specialization, metaphorical and metonymical changes as its recurring core) they are very much indebted to the then dominant theoretical frameworks (e.g., historical-philological semantics, structuralist semantics etc.) (Geeraerts, 2010). All taxonomies have been created using traditional, qualitative approaches and while relying on collections of texts, the approach are not data-driven.

The lack of resources tailored to train models on these complex semantic changes has prevented computational work thus far. Semantic change often entails a complete loss of previous meanings, making most contemporary computational lexical resources insufficient, as they typically miss the historical meanings. Although datasets exist for phenomena like metaphors, metonymies, or analogies, they often concentrate on the novel use of word meanings and omit conventional examples (Maudslay and Teufel, 2022) thus limiting the models' ability to predict semantic change types. While there exists no resources for training models, in this paper, we release a first benchmark for testing models on their ability to detect semantic change types given definitions.

3 LSC-CTD Benchmark

One of the most comprehensive and recent classification of semantic change is Blank's taxonomy (Blank, 1997). In this work, about 650 cases of semantic change are classified covering the vocabulary of the Romance languages and to some extent also German and English. Blank tries to reinterpret traditional types of classification from a cognitive-linguistics point of view. For the LSC Cause-Type-Definitions (LSC-CTD) Benchmark[4] we digitize the cases reported by Blank and create a benchmark of definitions manually curated by an historical linguist assisted by ChatGPT[5]. The format of the dataset, the dataset statistics and the annotation procedure are described further in Appendix A.

Blank's type classification include specialization, generalization, co-hyponymous transfer, auto-antonym, metaphor, antiphrasis, metonymy, auto-converse, ellipsis, folk-etymology, analogy, meaning dilution, meaning reinforcement, and doubtful cases. For our study, we use the following types of change for which we can find a counterpart in the synchronic lexical database.[6]

generalization: the old meaning is a sub-case of the new meaning (or: the new meaning has at least one less definitional feature than the old meaning); other terms used to refer to this type of change are *broadening* and *widening*. An example is the (non-attested) Vulgar Latin *adripare*, which originally meant 'to reach the shore (by vessel)' but in contemporary Romance languages has generalised to 'arrive

(by any means of transport)' (e.g.: French *arriver*, Italian *arrivare*)

specialization: the new meaning is a sub-case of the old meaning (or: the new meaning has at least one extra definitional feature than the old meaning); other terms used to refer to this type of change are *narrowing* and *restriction*. An example is Latin *necare*, which used to refer to 'to kill (by any means)' but whose French derivation *noyer* is now restricted to 'to kill by drowning'

co-hyponymous transfer: the change that occurs due to naming confusion, that is, when a word is used for new referents that are (wrongly) thought to be similar if not identical to the old referents. For instance, in some Romance languages the word for 'rat' is also used (in a derived form) to refer to 'mouse' or vice versa (Piedmontese: *rat* for 'rat' and *rató* for 'mouse'; Spanish: *ratón* for 'rat' and *rata* for 'mouse')

auto-antonymy: also called *contronymy* or *enantiosemy*, is when a word develops a new meaning that expresses a contrast, or is in opposition, to its old meaning. For instance, the Latin adjective *sacer* started out with the positive sense of 'sacred' and only later acquired the negative sense 'cursed', as evinced by the French *sacré* (Traugott, 2017; Karaman, 2008)

4 Generating examples from WordNet

To train a classifier, we use WordNet, a database that provides a large set of examples of lexical-semantic relations that are the synchronic counterpart of the semantic change types. WordNet is a multilingual synchronic lexical resource that can be queried to obtain semantic information for the covered vocabulary. Each *lemma* in WordNet is found within a *synset*, which brings together cognitive synonyms. Each synset therefore represent a sense or concept shared between several lemmas. Synsets are organized hierarchically in a tree structure, where the arc between a node and its parent represents a hyponymy relationship, meaning the lower node represents a more specific concept than that represented by the parent. This relationship can also be viewed inversely, where each parent node is connected to a child node if and only if there is a hyperonymy relationship between them,

that is, the parent expresses a more general concept than the child node.

WordNet associates each synset with a *sense gloss* and a series of *corpus examples*. The gloss functions as a concise definition which has been specifically conceived to uniquely express the meaning of the synset, and thus exemplify the latent construct of a sense. The examples, on the other hand, are mere uses of the words, which are often under-specified with respect to their meaning. For this reason, we focus on the sense gloss.

The relations of hyperonymy, hyponymy, co-hyponymy and antonymy can all be considered the synchronic counterpart of the semantic change types introduced in Section 3. Below we illustrate how we map the taxonomical relationship between the sense glosses of synsets in WordNet to each considered change type. Change types in Blank's taxonomy that cannot be modelled through the WordNet hierarchy are disregarded for this work. For hierarchical relationships of hyponymy, hyperonymy, and co-hyponymy, we consider paths of maximum length 1, meaning pairs for which there is a direct relationship.

Generalization and specialization: For every synset, we extract the list of associated hyperonymous synsets, and form pairs of a synset sense gloss with its hyperonym sense gloss. Viewed from the perspective of the child synset to the parent synset, we obtain instances of generalization. Similarly, by exploiting the inverse relationship (from parent to child synset), we obtain examples of specialization.

Co-hyponymy: We define a co-hyponymy relationship as two separate synset nodes with a common synset parent, in which the two child synsets represent two different and more specialized cases of the same general concept. By pairing the sense glosses of the child synsets, we generate the same sense relationship that lies at the basis of the diachronic co-hyponymous transfer.

Antonymy: WordNet also encodes the relationship between synsets that express opposite concepts, often between adjectives, but also found between other word classes such as nouns (*father* vs. *mother*) and verbs (*to buy* vs. *to sell*). For each synset, we therefore extract the synset, or the lemma within the synset, labeled as antonym, and pair the sense glosses of these two synsets.

Homonymy: Finally, we randomly generate pairs of synsets with the same part-of-speech that

do not lead to hyperonymous/hyponymous, co-hyponynic and antonymic pairings of sense glosses. The potential unrelatedness between the senses should be a good proxy for homonymy, defined as the occurrence of completely unrelated senses (for the same word form). While it is possible that related senses of the same word are paired in the random pairing, the risk is extremely small (i.e., on average 3-4 out of the 30-40,000 randomly sampled synsets can belong to the same word). Homonymy is not treated on the same par as the other types of semantic change by Blank (1997), even though its importance as a driver of language change is uncontroversial. For this reason it has been included in the experiments.

Semantic Change type	Synchronic relation
Generalization	Hyperonymy
Specialization	Hyponymy
Co-hyponymous transfer	Co-hyponymy
Auto-antonymy	Antonymy
Unrelated	Homonymy

Table 2: Mapping between Semantic change types to the synchronic relations found in WordNet.

The WordNet testset: We divide the final dataset into three distinct datasets: training, validation, and testing, selecting 80%, 10%, and 10% of the data respectively. The examples are split in such a way that the same pair of definitions for a specific class never appears simultaneously in two datasets (training/validation, validation/test, training/test). Furthermore, to avoid class imbalance, we limit the maximum number of examples per class to 30,000, 3,000, and 3,000 for training, validation, and testing respectively. The statistic of the dataset are documented in Appendix C.

5 Model

To classify the relationship between senses, we utilize the data collected in Section 4 and train a classification model[7] that takes as input the two definitions of the respective senses and outputs the corresponding class, i.e. {*homonymy, hyperonymy, hyponymy, antonymy*}.

Specifically, we concatenate the textual definitions δ_1 and δ_2 together with special tokens into a

---

[7]Our code is available on GitHub https://github.com/ChangeIsKey/change-type-classification. The classification model is available on the Hugging Face Model Hub https://huggingface.co/ChangeIsKey/change-type-classifier.

single string $\delta = \texttt{<s>}\delta_1 + \delta_2\texttt{</s>}$, which is then tokenized and encoded using roberta-large (Liu et al., 2019). From this, we obtain the vectors corresponding to the sub-tokens of the original string $v_1, .., v_K$. The vector v_1, corresponding to the initial token, is used for classification. We apply a linear transformation to reduce the vector's dimension to the label space, such that $v_1^* = W v_1$, where $W \in \mathbb{R}^{C \times N}$ and C is the number of classes. Finally, we use cross-entropy to optimize the model.

To counteract class imbalance, examples in each batch are drawn proportionally from each class. We use AdamW for optimization with a weight decay of 0.1 and a learning rate of 1e-6. Early stopping is employed to determine when to halt training based on accuracy on the development set. Additionally, we assess the outcomes using two baselines. The results for these baselines are detailed in Appendix B.

6 Evaluation of the Type Classification

We first evaluate the model on the test portion of the synchronic relations found in WordNet. Next, we test the model on a historical, novel data, namely the LSC-CTD benchmark.

Figure 1: Confusion matrix for predictions of sense relationships on the WordNet testset.

6.1 Evaluation on WordNet

We evaluate the classifier on the WordNet testset and investigate each class individually using the confusion matrix found in Figure 1. The diagonal results show correctly classified predictions for each class. The normalized recall rate range from

4543

0.83 for the co-hyponym class to 0.95 for the hono-mym class. That means, the model correctly identi-fies between 83% – 95% of the true class members showing that the model performs well in distin-guishing these classes.

For hyperonyms and hyponyms, the model shows a normalized recall rate of 0.91 and 0.90 respectively. The confusion mainly occurs with the converse relationship (hyponyms/hyperonyms) (0.048/0.057), which is logical given the close re-lationship between hyperonyms and hyponyms, as a hyperonym's definition is inherently included within its hyponym's definition. There are minor confusions with co-hyponyms (0.03) and a negligi-ble amount with antonyms and homonyms.

The model exhibits the lowest normalized recall rate of 0.83 for co-hyponyms. Despite being al-most 10-points lower than that for hyperonyms and hyponyms, it still suggests a substantial capabil-ity to correctly classify co-hyponyms. The con-fusion is distributed across the other hierarchical relationships. Co-hyponyms get misclassified as hyperonyms (0.07), hyponyms (0.06), and to a very small proportion as homonyms (0.01). The slightly lower performance in this category could be due to the more complex nature of co-hyponym rela-tionships, which require the model to understand not just a direct hierarchical link, but also parallel connections within the same level of hierarchy.

The negative class serves as a control or base-line class against which hierarchical or antonym relationships are compared. The model correctly identify 95% of instances that do not exhibit a hier-archical relationship.

The small fractions of misclassification with other classes (0.008 with hyponym, 0.009 with hy-peronym, 0.026 with co-hyponym, and 0.004 with antonym) indicate that the model is very effective at distinguishing non-hierarchical and non-antonym relations from homonyms ones. This high perfor-mance is crucial as it shows the model's strong dis-criminatory power, effectively reducing the false positive rate where hierarchical or antonym rela-tionships are incorrectly inferred.

6.2 Evaluation on the LSC-CTD Benchmark

To answer how well the model trained on syn-chronic relations can classify diachronic relations, i.e., semantic change types, we test the model on the LSC-CTD Benchmark. We follow the mapping shown in Table 2 and consider a prediction correct within the mapped class. As an example, a pair of

definitions classified as having exhibited special-ization is correct if the model predicts hyponomy. Result are shown in Figure 2.

Figure 2: Confusion matrix for predictions of sense relationships on the LSC-CTD Benchmark.

The model shows a high normalized recall rate of 0.85 for Specialization, that is, 85% of all the spe-cialization instances are correctly predicted as hy-ponyms. The confusion occurs with co-hyponyms (0.15). This suggests that while the model is gen-erally good at recognizing the specialization that occurs from a general to a more specific sense, it can sometimes fail to see the vertical relationships between these senses and consider them on the same horizontal level (as co-hyponyms).

The model exhibits a normalized recall rate of 0.68 for generalization, i.e., moderate success in identifying instances of generalization. There is a notable amount of confusion with hyponyms (0.16), and co-hyponyms (0.11). This asymmetry between generalization and specialization classes could be due to the fact that going from more specific to more general definitions often implies the use of fewer and more abstract or polysemous terms, which are notoriously harder to represent.

The model has a low normalized recall rate of 0.17 for co-hyponymous transfer, suggesting sig-nificant challenges in accurately classifying this type of semantic change. The confusion is well distributed across the other classes. with the largest misclassification being with hyponyms (0.46) and hyperonyms (0.29). There might be two reasons for this: co-hyponymous transfer fundamentally relies on confusion and referential ambiguity on the side of the speaker, so it is not surprising that specifically text-based LLMs cannot handle such confusion; in Blank's dataset it is considered a phe-nomenon of rather limited range, mainly observed
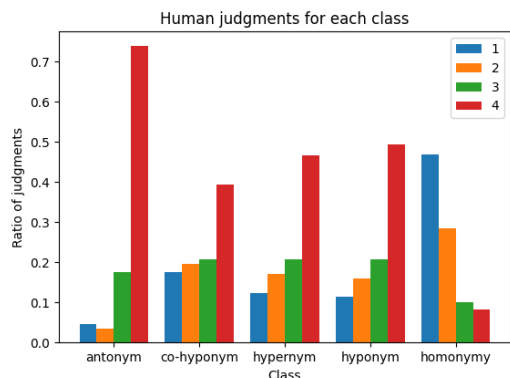
Figure 3: Ratio of human judgments for each class.

in specific dialects (a third of the examples come from Réunion Creole). The model seems to have a moderately good grasp of auto-antonymy correctly classifying 62% of these instances. Here there is room for improvement, in particular by increasing the number of training examples on synchronic antonym relations.

7 Evaluation on Semantic Change

We will use our classification of semantic change type, introduced in Section 5, to test if we can improve on state-of-the-art LSC detection. In this section, we use Flan-T5 XL FT presented by Giulianelli et al. (2023) to generate definitions, which is a Flan-T5 XL (Chung et al., 2022) model finetuned on WordNet (Miller, 2013), the Oxford dataset (Gadetsky et al., 2018) and the CoDWoE dataset (Mickus et al., 2022).

7.1 Semantic Relatedness and Homonymy

In this Section, we use the English portion of the SemEval-2020 Task 1 to study the correlation between human annotated relatedness scores and word sense relationships.

In the annotation of SemEval-2020 Task 1, (see further Section 2.1), pairs of usages are sampled for each word. These usage pairs are either synchronic, i.e. belong to the same time period, or diachronic, i.e. belong to different time periods. For each pair of usages, the annotators have to assign a label on a scale from 1 to 4 (Table 1): 1 (unrelated), 2 (distantly related), 3 (closely related), and 4 (identical). This scale draws inspiration from the work of Blank (1997), which introduced the concept of a continuum of semantic proximity. Specifically, Schlechtweg et al. (2018) provides a direct mapping between the relatedness scale and the semantic

proximity introduced by Blank (1997).

For each usage in the English portion of the SemEval-2020 Task 1, we generate a definition. We then predict word sense relationships using the definitions for all the usage pairs for which human judgments are available. The word sense relationships include homonymy, co-hyponymy (co-hyponymous transfer) , hyperonymy (generalization), hyponymy (specialization) and antonymy (auto-antonymy). The bar graph in Figure 3 represents, for each predicted class, the distribution of the existing human judgments based on the semantic relatedness scale[8].

The homonymy category has a high number of judgments in class 1 (Unrelated) (60% of the overall unrelated pairs), which is consistent with the definition of homonyms: words that are spelled the same and sound the same but have different meanings. There are very few 3 or 4 judgments, which suggests that humans generally do not find homonyms to be semantically related.

Overall, the distribution of human judgments reflects an understanding of semantic relationships that aligns with the theoretical definitions of each relationship type. Closely related judgments dominate in categories where a semantic link is expected (co-hyponym, hyperonym, hyponym), while unrelated judgments are more frequent for homonymy samples, where no semantic relationship or an opposite relationship is anticipated.

7.2 Graded WiC Task

The model's exceptional ability to identify unrelated pairs annotated with 1 clearly demonstrates its adequacy in modeling the homonymy class. We will use this ability to improve models for the Graded WiC task.

We set up the evaluation as a Graded WiC task, directly comparing the semantic proximity value provided by human annotators and the label computed by the model. To do this, we transform the model's computed labels into a two-value scale: 1 (Related) and 0 (Unrelated), grouping the labels hyperonym, hyponym, co-hyponym, and antonym under the Related class, while the homonymy class represents the Unrelated class.

Furthermore, we investigate the combination of XL-LEXEME and Definitions + Homonym using

---

[8]It is important to note that the type classification model used in the study is not designed to identify identical meanings, which corresponds to a human judgment of 4.

Model	Correlation
Definitions + SacreBLEU	0.108
Definitions + METEOR	0.117
Definitions + Cosine similarity	0.264
Definitions + Homonym	0.472
XL-LEXEME	0.623
Definitions + Homonym + XL-LEXEME	**0.646**

Table 3: Spearman correlation of human judgments vs model predictions, Definitions generated using the method of Giulianelli et al. (2023)

a weighting scheme:

$$\rho(u_1, u_2) = \begin{cases} \cos(u_1, u_2), & \text{if } u_1, u_2 \ Related. \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

For the evaluation, we calculate the Spearman correlation between human judgments and the labels computed as described. The results are presented in Table 3, where we compare our model with XL-LEXEME, which represents the state of the art in this task, and other strategies for calculating the distance between definitions proposed by Giulianelli et al. (2023). As the results show, the classification model achieves significantly better results compared to using metrics like BLEU or cosine distance applied to definition vectors calculated using SBERT. On the other hand, the model performs well below the state of the art, with a 15-point difference from XL-LEXEME. However, when we combine the classification model with XL-LEXEME, we achieve a new state-of-the-art result, indicating that the classification model produces reliable labels when classifying homonymy, i.e., instances annotated with 1.

7.3 Binary Change Task

In SemEval-2020 Task 1, and in general for DWUGs datasets, words are annotated as stable or changed based on the removal and/or addition of a meaning (i.e., cluster). In this process, annotators' judgments are binarized, so that pairs of usages annotated with a score lower than 2.5 represent negative edges, while the rest of the usages represent positive edges. The optimization algorithm then seeks to group positive edges into clusters, keeping them separate from negative ones, with the resulting clusters representing word senses.

Given the high correlation of the homonymy class with negative judgments (1 and 2), this represents an ideal proxy for automatically classifying words that have changed in meaning. Specifically,

Model	Accuracy
Definitions + Homonym	**0.783**
XL-LEXEME + 0.5 threshold	0.761
XL-LEXEME + Opt. threshold	0.848

Table 4: Binary task SemEval-2020 Task 1 (EN)

to classify words that have changed meaning in the SemEval dataset, we count for each word the occurrences of each label returned by the classification model, specifically the counts of hyperonym, hyponym, co-hyponym, antonym, and homonymy. We assign the label 1 (changed) to those words for which the homonymy class is the most frequent.

Table 4 shows the accuracy on SemEval-2020 Task 1 - Subtask 1 for the English language. The results show that the Definitions + Homonym classification model achieved 78.3% accuracy without the need for threshold adjustment. The XL-LEXEME model reached 76.1% accuracy using the threshold of 0.5 (with which it has been trained), while optimizing the threshold improved its accuracy significantly to 84.8%. It is notable that the Definitions + Homonym classification model performs well without the need to set up a threshold, which indicate that it is robust and less sensitive to parameter tuning compared to the XL-LEXEME model.

8 Conclusion

With this paper, we present a first computational approach to detecting the type of semantic change that a word has experienced. We release the LSC-CTD Benchmark, a first dataset for computational modeling of semantic change that is annotated by type of change. The benchmark is based on a digitized version of Blank's taxonomy extended with concise sense definitions. Because the resource is limited in size, it can only be used for testing purposes. To train a model for classifying change types, we used WordNet with synchronic sense definitions and relations between the synsets.

The use of WordNet as a foundational resource for training our classifier enabled us to leverage a vast repository of semantic relationships and meanings. Our evaluation is grounded both in linguistic theory and computational metrics. We prove that homonyms are perceived by human annotators as the most distant semantic difference class in Blank's continuum of semantic proximity.

We applied our classifier to extend semantic change detection models with type information and tested it on a standard evaluation benchmark,

namely SemEval-2020 Task 1. We aimed to provide a more refined and accurate model for semantic change detection. We found that by incorporating change types, in particular the notion of homonymy, we improved the state-of-the-art results for LSC on the graded task. For the binary task, we also saw encouraging results.

Our findings underscore the importance of distinguishing between different types of semantic changes, such as generalization, specialization, and co-hyponymous transfer, in understanding the dynamics of semantic change. In future work, we will extend our resources with more examples of change types, as well as additional synchronic resources.

9 Limitations

Despite the promising outcomes, we acknowledge limitations in our study, regarding the employed resource, the limits of the distributional hypothesis and our focus on lexical type instead of token.

First, Blank's database is a rich lexical resource, unique in its sort, but is nevertheless the work of one scholar working from one theoretical background. This likely introduces bias in the selection and diversity of the examples. Another technical limitation of the benchmark is that we we took the generated definitions for granted as long as they were factually correct, but more detailed control on the specific wordings (for instance: to avoid specific lexical cues that bias the model in one or the other direction) might improve the results in the future.

Second, from a theoretical point of view, the types of semantic change under scrutiny do not emerge from text, but are the result of (oral) communicative exchanges and cognitive processes, and how both of these change over time. In other words, written input/context will always fall short of providing semantic information that is sufficiently rich to obtain optimal sense representations. Our choice of working with definitions partially circumvents this problem of under-determinism, but in general the limitation pertains to all kinds of context-based methods relying on the distributional hypothesis. As such, also our generated definitions are probably affected by it.

Third, our initial approach to tackling the problem starts from lexical *types* and their definition, for reasons of control and feasibility, but, as discussed above, semantic change unfolds in historical contexts and were likely, at the time of creation, selected from concrete communicative situations.

The most appropriate level for modeling these phenomena is therefore at the level of lexical *tokens*. To test this in the wild, we need to have manually annotated, historical resources in which change types are marked. This will enable studies to go from the level of definition to the level of usages. Our team is working on constructing such an annotated resource, but this is both expensive and time consuming and will likely only be at such a level that it can be used for testing purposes.

Acknowledgments

References

Nikolay Arefyev, Daniil Homskiy, Maksim Fedoseev, Adis Davletov, Vitaly Protasov, and Alexander Panchenko. 2021. DeepMistake: Which Senses are Hard to Distinguish for a WordinContext Model. In *Computational Linguistics and Intellectual Technologies - Papers from the Annual International Conference "Dialogue" 2021*, volume 2021-June. Section: 20.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, Online. Association for Computational Linguistics.

Andreas Blank. 1997. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*, volume 285 of *Beihefte zur Zeitschrift für romanische Philologie*. Niemeyer, Tübingen.

Leonard Bloomfield. 1933. *Language*. Holt, Rinehart, and Winston, New York.

Michel Bréal. 1904. *Essai de sémantique (science des significations)*. Hachette.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Arsène Darmesteter. 1893. *La vie des mots étudiée dans leurs significations*. C. Delagrave.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 266–271, Melbourne, Australia. Association for Computational Linguistics.

Noah Gardner, Hafiz Khan, and Chih-Cheng Hung. 2022. Definition modeling: literature review and dataset analysis. *Applied Computing and Intelligence*, 2(1):83–98.

Dirk Geeraerts. 2010. *Theories of Lexical Semantics*. Oxford University Press, Oxford.

Dirk Geeraerts. 2020. Semantic change. "what the smurf?". In Daniel Gutzmann, Lisa Matthewson, Cécile Meier, Hotze Rullmann, and Thomas E. Zimmermann, editors, *The Wiley Blackwell Companion to Semantics*. Wiley Blackwell, Hoboken NJ.

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.

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.

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. Definition modelling for appropriate specificity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2499–2509, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Burcu I. Karaman. 2008. On Contronymy1. *International Journal of Lexicography*, 21(2):173–192.

Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and improving BERT performance on lexical semantic change detection. In *Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, EACL 2021, Online, April 19-23, 2021*, pages 192–202. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arxiv*, abs/1907.11692.

Rowan Hall Maudslay and Simone Teufel. 2022. Metaphorical polysemy detection: Conventional metaphor meets word sense disambiguation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 65–77, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Timothee Mickus, Kees van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 task 1: CODWOE - comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022*, pages 1–14. Association for Computational Linguistics.

Justin J Miller. 2013. Graph database applications and concepts with neo4j. In *Proceedings of the southern association for information systems conference, Atlanta, GA, USA*, volume 2324.

Stefano Montanelli and Francesco Periti. 2023. A Survey on Contextualised Semantic Shift Detection.

Roberto Navigli. 2009. Word Sense Disambiguation: A Survey. *ACM Comput. Surv.*, 41(2).

Hermann Paul. 1880. *Prinzipien der Sprachgeschichte*. Niemeyer, Halle.

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*, Mexico City, Mexico. Association for Computational Linguistics.

Maxim Rachinskiy and Nikolay Arefyev. 2021. Zeroshot Crosslingual Transfer of a Gloss Language Model for Semantic Change Detection. In *Computational Linguistics and Intellectual Technologies - Papers from the Annual International Conference "Dialogue" 2021*, volume 2021-June. Section: 20.

Karl Christian Reisig. 1839. *Professor K. Reisig's Vorlesungen über lateinische Sprachwissenschaft*. Verlag der Lehnhold'schen Buchhandlung.

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.

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.

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.

Gustaf Stern. 1931. *Meaning and change of meaning; with special reference to the English language.* Indiana University Press, Bloomington.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. *Survey of Computational Approaches to Lexical Semantic Change Detection*, pages 1–91. Language Science Press, Berlin.

Elizabeth Closs Traugott. 2017. Semantic change. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

S. Ullmann. 1957. *The Principles of Semantics*. Glasgow University publications. Jackson.

	Train	Development	Test
Unique synsets	79469	15994	15978
Definition pairs	123034	12379	12380
Hyponyms pairs	30000	3000	3000
Hyperonyms pairs	30000	3000	3000
Co-hyponyms pairs	30000	3000	3000
Homonyms pairs	30000	3000	3000
Antonyms pairs	3033	379	380
Noun pairs	85084	8535	8480
Verb pairs	20772	2032	2137
Adjective pairs	14268	1491	1473
Adverb pairs	2910	321	290

Table 5: WN statistics.

A LSC-CTD Benchmark

In Table 6, we present an excerpt from the LSC-CTD Benchmark, a manually curated dataset of dictionary-like definition pairs for words that experienced semantic change. It includes Blank's original annotations detailing the target word followed by a language tag (lt: Latin, vlt: Vulgar Latin) and German glosses of the original, primary meaning-(Old Meaning) and of the derived, secondary meaning (New Meaning). We translate Blank's glosses into English using Google Translate's API (between double quotation marks). Following this, we use the prompt shown in Figure 7 to generate definitions that are more akin to those found in a dictionary, employing gpt-4 API (between single quotation marks). This prompt includes some examples we initially set up, with the actual generation occurring through few-shot learning. Furthermore Blank's dataset contains the alleged cause of the semantic change and, not shown here, the higher-level association relation that exists at the conceptual level. Finally, a linguist with expertise in historical linguistics and semantics manually reviewed and refined the collection of generated translations and definitions, replacing or modifying them as necessary.

B Baseline

As our baselines, we consider two models. The first is naturally the base version of roberta (roberta-base) (Liu et al., 2019), as we have utilized the large version for the experiments reported in the paper (see Figure 5). The training of the base

model is identical to that of the large model, employing the same hyperparameters and data. The second baseline, on the other hand, is based on the tf-idf weighting model (see Figure 4). Specifically, we calculate weights across all definitions present in the training set. Subsequently, we extract vectors for each definition and use the concatenation of vectors from pairs of definitions as features for our model. In particular, we employ a simple classifier based on stochastic gradient descent (SGD), utilizing the implementation provided by scikit-learn[9] with default parameters, except for class weighting where we attempt to mitigate imbalance using the balanced parameter. The tf-idf baseline is used to investigate trivial patterns among pairs of definitions, that is, whether the simple co-occurrence of words can be a signal for identifying the target class. Not many differences are observed between roberta-base and roberta-large, both on the WN test set and the LSC-CTD benchmark. In roberta-large there is however a relatively large improvement in the detection of auto-antonymy (+ 0.12) and generalization (+ 0.03), but also a lower performance for co-hyponymous transfer (-0.04). The comparison between the roberta models and the tf-idf models is more striking, with the latter model failing on differentiating homonymy from auto-antonymy (probably due to not taking into account the syntactic position of negations) again on both test sets. Surprisingly, regarding the identification of co-hyponymous transfer, the tf-idf model performs much better (+0.16 relative to roberta-large).
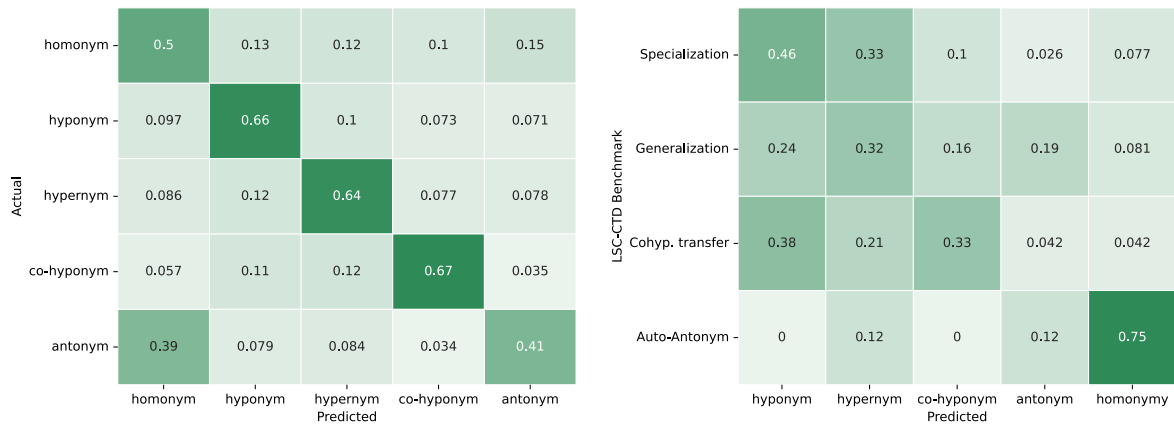
C WN Statistics

In Table 5 we report the statistics we use to train the sense relationship model. The table includes for each split (train, dev, and test set) the unique number of synsets, the number of pairs of definitions, and the number of pairs of definitions for each class (hyponym, hyperonym, co-hyponym, homonym, and antonym). Additionally, the number of distinct pairs for each part of speech (nouns, verbs, adjectives, and adverbs) is reported.

[9] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html

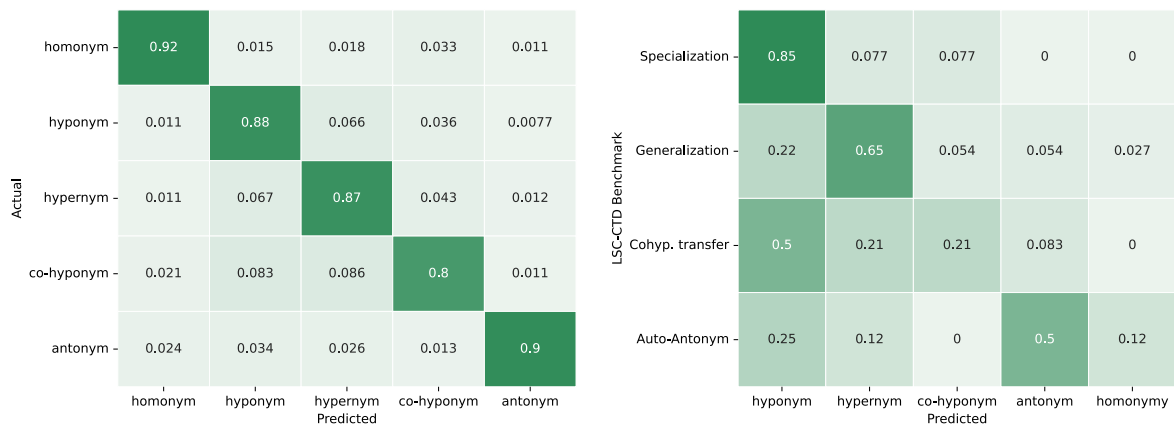| Word	Old Meaning	New meaning	Cause	Type
adripare:vlt	am Ufer ankommen "arrive at the bank/shore" 'arrive at the bank of a river or the shore of a lake or sea'	ankommen "arrive" 'to reach a place, especially at the end of a journey'	prototype / frame	generalization
necare:lt	töten "kill" 'to cause the death of a living thing, typically involving an act of violence or an intention to harm.'	ertränken "drown" 'to cause to die by submersion in liquid, especially by forcing the head under the water.'	socio-cultural change	specialization
ratta	Ratte "rat" 'a small rodent, larger than a mouse, that has a long tail and is considered to be harmful'	Maus "mouse" 'a small mammal with short fur, a pointed face, and a long tail'	referential vagueness	co-hyponymous transfer
sacer:lt	heilig , geheiligt "sacred" 'considered to be holy and deserving respect, especially because of a connection with a god'	verflucht "cursed" 'experiencing bad luck caused by a magic curse'	taboo	auto-antonymy

Table 6: Snippet of the LSC Cause-Type-Definitions Benchmark

(a) WN test set.

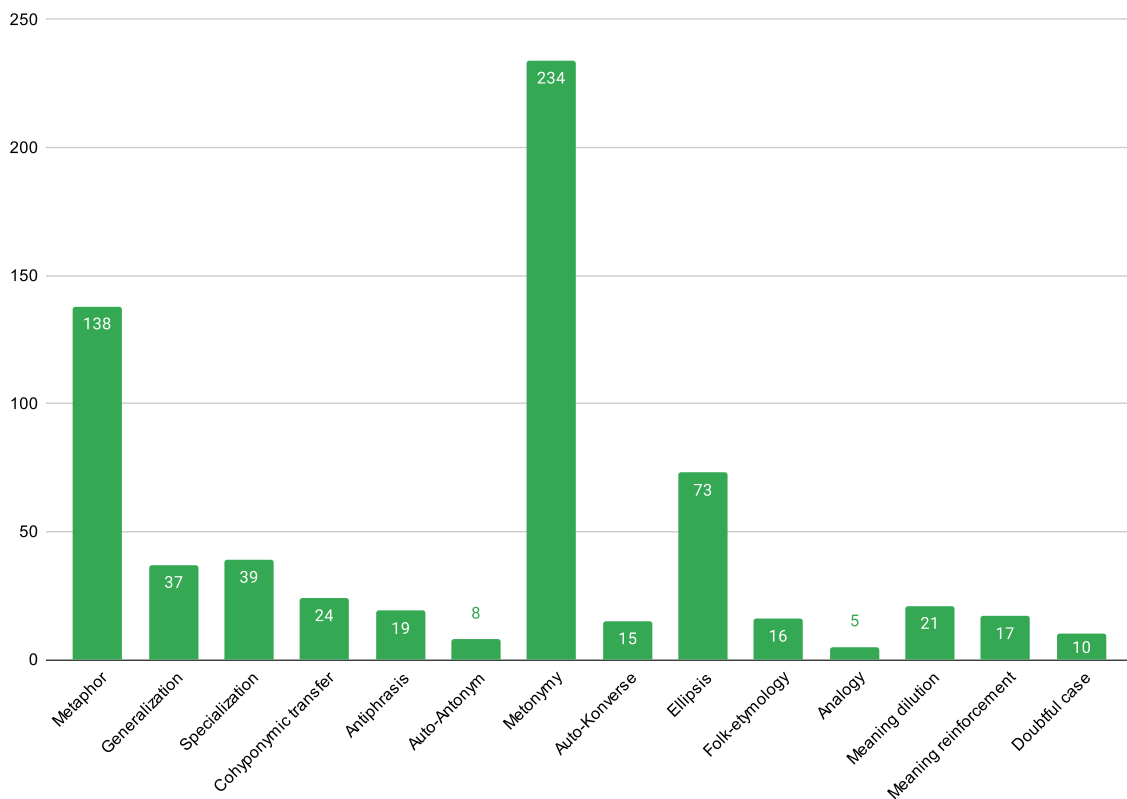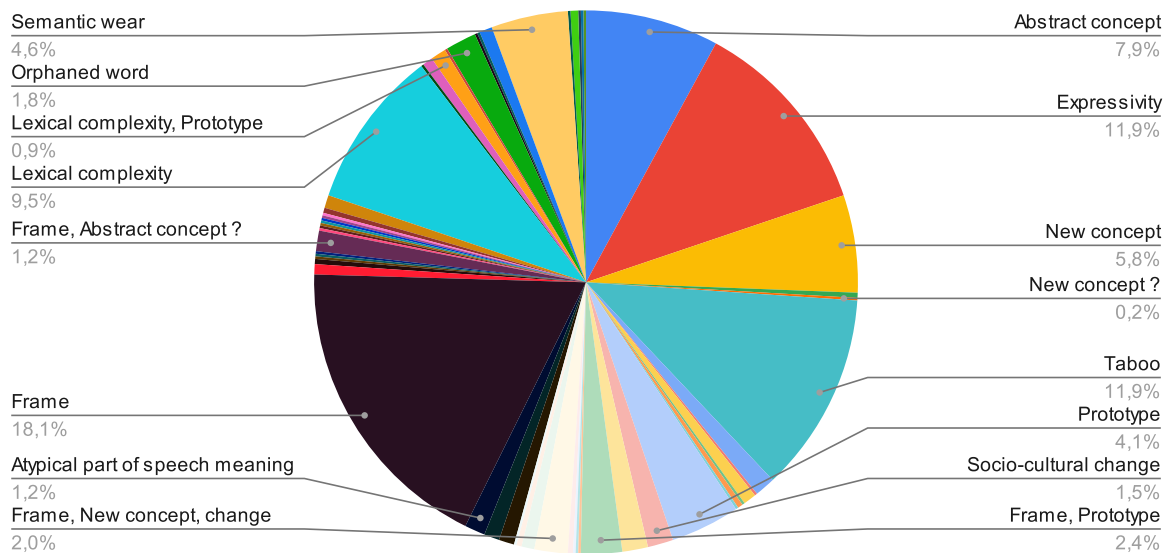(b) LSC-CTD Benchmark.

Figure 4: SGD TF-IDF.

(a) WN test set.

(b) LSC-CTD Benchmark.

Figure 5: roberta-base.

(a) Number of cases for each type.

(b) Number of cases for each cause.

Figure 6: LSC CTD Benchmark statistics.

Words	Old Meaning	New Meaning	Cause	Association Model	Notes	Type	Old Meaning (EN)	New Meaning (EN)	Old Meaning (EN Definition)	New Meaning (EN Definition)
afferrare:it	ergreifen	verstehen	abstraktes Konzept	SimiDES	KörperGeist	Metaphor	grab	understand	to quickly take…	to understand …
affreusement:fr	scheußlich	sehr, stark (Adv.)	Expressivität	SimiDES	Intensität	Metaphor	awful	very, strong (adv.)	extremely bad…	very much or in a…
agenda:sp	Notizbuch	NotebookComputer	neues Konzept	SimiDES	Gegenstand Gegenstand	Metaphor	Notebook	Notebook computer	a book of paper for…	a small, light computer
águila:sp	Adler	schlauer Mensch, Fuchs	Expressivität	SimiDES	TierMensch	Metaphor	Adler	Clever person, fox	a large, strong bird…	having or showing…
aile:fr; ala:it,sp,pt	Flügel	Flügel eines Gebäudes	neues Konzept	SimiDES	Tier Gegenstand	Metaphor	wing	Wing of a building	the flat part of the…	a part of a large…

Generate Old Meaning (EN Definition) and New Meaning (EN Definition) for the following entries, providing descriptive, dictionary-like definitions as in the examples above.

Figure 7: Chat GPT prompt.