Sense Through Time: Diachronic word sense annotations for Word Sense Induction and Lexical Semantic Change Detection

Dominik Schlechtweg[1†], Frank D. Zamora-Reina[2†], Felipe Bravo-Marquez[2], Nikolay Arefyev[3]

[1*]Institute for Natural Language Processing, University of Stuttgart.
[2]Department of Computer Science, University of Chile, CENIA & IMFD.
[3]Department of Informatics, University of Oslo.

Contributing authors: schlecdk@ims.uni-stuttgart.de; fzamora.reina@gmail.com; fbravo@dcc.uchile.cl; nikolare@uio.no;
[†]These authors contributed equally to this work.

Abstract

There has been extensive work on human word sense annotation, i.e., manually labeling word uses in natural texts according to their senses. Such labels were primarily created for the tasks of Word Sense Disambiguation (WSD) and Word Sense Induction (WSI). However, almost all datasets annotated with word senses are synchronic datasets, i.e., contain texts created in a relatively short period of time and often do not provide the creation date of the texts. This ignores possible applications in diachronic-historic settings, where the aim is to induce or disambiguate historical word senses or changes in senses across time. To facilitate investigations into historical WSD and WSI and to establish connections with the task of Lexical Semantic Change Detection (LSCD), there is a crucial need for historical word sense-annotated data. Hence, we created a new reliable diachronic WSD/WSI dataset 'DWUG DE Sense'. We describe the preparation and annotation and analyze central statistics. We then describe a thorough evaluation of different prediction systems for jointly solving both WSI and LSCD tasks. All our systems are based on a state-of-the-art architecture that combines Word-in-Context models and graph clustering techniques with different hyperparameter settings. Our findings reveal that using the WSI task as optimization criterion yields better results for both tasks even when the LSCD task is the focal point of optimization. This underscores that although both tasks are related, WSI seems to be more general and able to incorporate the LSCD task.

Keywords: Word Sense Induction, Word Sense Disambiguation, Lexical Semantic Change Detection, Word-in-Context, historical, diachronic

1 Introduction

There has been extensive work on manual annotation of word occurrences in natural texts with word senses [e.g. 1–3] following from an early interest of computational linguists in the tasks of Word Sense Disambiguation [WSD, 4] (i.e., the task of disambiguating an ambiguous word given its context) and Word Sense Induction [WSI, 5] (i.e., the task of inducing word senses without predefined sense inventory). However, the work on (traditional) word sense annotation has been done nearly exclusively with a **synchronic** focus, i.e., annotated word uses have been sampled from narrow time spans in modern corpora. Consequently, the tasks defined on this data and models developed to solve them also have a synchronic focus. This ignores possible applications of WSI and WSD in historical settings, where the aim is to induce or disambiguate historical word senses or changes in senses across time [6, 7]. Solving these tasks can be helpful to create historical or etymological dictionaries [e.g. 8, 9] or inform linguistic analysis [e.g. 10, 11]

In recent years, more and more models solving tasks of lexical semantics have been applied to **diachronic-historic** data, i.e., word uses sampled from various time spans in historical corpora [12, 13] with the main aim of detecting changes in meaning of words over time, known as the task of Lexical Semantic Change Detection [LSCD, 13]. There are several LSCD models with standard WSD [e.g. 14, 15] or WSI [e.g. 16–18] components [find a recent overview in 19]. Although these components are central to the models, our experience with the behavior of WSD and WSI models on diachronic-historic data remains limited. There are indeed specific challenges in WSI on diachronic-historic data is demonstrated by Laicher et al. [17] showing that standard clustering algorithms with off-the-shelf BERT [20] are very sensitive to historic spelling variations.

Additionally, the LSCD task has been formulated such that it does not explicitly require systems to assign word uses (drawn from two distinct time periods) to their corresponding senses as the primary evaluation metric focuses on how effectively these systems quantify the extent of change for some given words [cf. 13]. This is also due to the fact that previous LSCD datasets were not sense-annotated. However, going beyond the mere detection of change would be extremely valuable for historical linguistic research. For example, by asking systems to qualify senses which were lost or gained.

We argue that in order to bridge the gap between these lines of research, we need diachronic-historic word sense-annotated data. Hence, we created a new, large and reliable WSI dataset which we call 'DWUG DE Sense', based on the existing LSCD dataset DWUG DE [13] and conducted various experiments on WSI and LSCD tasks with our data.[1] To the best of our knowledge, this is the first German dataset providing

[1] The dataset has been introduced originally in Schlechtweg [21, pp. 57–58], where it has been used to validate a second use-use annotated dataset. We go beyond this previous work by adding a much more

word sense annotations for different time periods. More specifically, we aim to address the following research questions:

RQ1. Can human annotators achieve a similar agreement on historical data as on modern data in use-sense annotation?
RQ2. Does WSI performance vary between periods?
RQ3. Can a single system be optimized to effectively solve both WSI and LSCD for diachronic data?

In this article, we address these questions through comprehensive experimentation and analysis. First, we describe the preparation and annotation of the dataset and analyze the agreement between annotators from a synchronic and diachronic viewpoint. Second, we conduct a comprehensive evaluation of several systems using a state-of-the-art pipeline [22] capable of simultaneously addressing WSI and LSCD on our dataset. These systems use a Word-in-Context (WiC) model to compute the probability that two word uses have the same sense, which is used to build a weighted graph for each word with all probabilities between the word's uses. Finally, a graph clustering algorithm is used to generate sense clusters, allowing predictions for both WSI and LSCD tasks. Our experiments mainly involve the manipulation of the WiC model, the graph clustering method, and several other hyperparameters. The main evaluation focus is on transportability of optimized parameters between WSI and LSCD in order to answer RQ3. We hypothesize that optimizing parameters for one of the tasks, also yields optimal parameters for the other one.

This paper is divided into seven sections. Section 2 provides a comprehensive overview of the existing research on WSI and LSCD, along with key definitions relevant to these fields. In the following section, we describe the creation of our dataset, exploring its structure, agreement, and the aggregated annotations. Section 4 introduces the two tasks addressed (WSI and LSCD), taking into account the specific characteristics of the constructed dataset. Section 5 describes the models used in this study. Section 6 presents the experiments and discusses the results. Finally, Section 7 presents our main conclusions and directions for future research.

2 Related work

2.1 Word meaning annotation

Throughout the paper, we will mean by a **word use** an occurrence of a word within an instance of text such as a sentence or a paragraph. By a **sense definition** we will mean a textual description of some meaning a target word can express. For shortness, we will sometimes refer to these concepts by 'use' and 'sense' if the context sufficiently disambiguates what we mean.

---

detailed analysis of the data including annotation examples (Section 3.1.1), statistics (Table 1), time-specific agreement analysis (Table 2), aggregated data examples (Figure 2 and 3 with discussion) and by performing WSI and LSCD experiments with computational models (Sections 4–6).

Word meaning annotated data falls into three main categories: (i) use-sense, (ii) use-use and (iii) lexical substitution annotation [cf. 3, 21, p. 20]. The first type, use-sense annotation has a long tradition within the task of WSD [4]. Annotators usually choose the best-fitting word sense definition for a word use as in this example:

> **use**: [...] and taking a knife from her pocket, she opened a vein in her little **arm**, and dipping a feather in the blood, wrote something on a piece of white cloth, which was spread before her.
> **sense1**: a human limb
> **sense2**: weapon system

There has been extensive work on use-sense annotation and several large-scale annotation projects have been carried out, as e.g. SemCor and OntoNotes [1, 2]. Such data can also be used for the WSI task as it provides a mapping of uses to clusters, i.e., all uses annotated with the same sense definition receive the same cluster label [cf. 5, 23].

The (discrete) use-sense assignment approach was criticized to be empirically inadequate [24–27]. This was (amongst others) supported by the observation that inter-annotator agreement for certain words was consistently low [2, 28]. Hence, the alternative annotation strategies for use-use pairs and lexical substitutions were developed. In the former approach annotators typically judge uses of a word for their semantic proximity [3, 29] while in the latter approach annotators find synonymous substitution words [30]. However, these alternative strategies do not solve the problem of low agreement [cf. 3, 30] and they do not directly provide data usable for WSD and WSI [31]. While word sense clusters for WSI can be obtained through clustering algorithms McCarthy et al. [31], Schlechtweg et al. [32, 33], such approaches are currently not thoroughly validated.

There is a growing body of work with a diachronic focus on word meaning [e.g. 7, 12, 34], tackling tasks such as LSCD [21]. This development has brought a number of diachronic word meaning datasets [e.g. 16, 19, 29, 32, 35, 36]. In contrast to their synchronic counterparts, these datasets are mainly annotated within the use-use paradigm.[2] On several of these datasets word sense clusters have been inferred from the human annotation [e.g. 19, 32, 37, 38] which could be used for WSI tasks in a diachronic setting. However, the inferred clusters often suffer from ambiguity and sparsity of annotation [21, p. 57ff.]. While we deem this a promising approach, it is still ongoing research.

Despite the above-mentioned criticisms, the traditional use-sense annotation approach has several advantages over the use-use setting: (i) The number of annotations needed is much lower [21, p. 45]. (ii) The sense clusters follow directly, so no additional sense/cluster inference procedure introducing additional noise is needed. (iii) The annotation is richer as it provides a sense definition and thus allows to define more tasks on the data, as e.g. WSD. (iv) The agreement is higher than with other approaches [cf. 3, 30]. (v) The data can be cleaned easily as agreement is calculated on instances rather than pairs of instances, hence individual instances can easily be excluded. However, to our knowledge only a small fraction of diachronic datasets is

[2] This is probably an influence from the early DURel annotation approach [29] relying on the insight that one does not have to annotate word senses in order to annotate (specific types of) word sense change.

use-sense annotated [39–42]. Of these only McGillivray [41]'s Latin is publicly available. The sense example sentences from historical dictionaries [e.g. 8, 9, 43] can be seen as use-sense annotated data [cf. 44]. However, these sentences do not provide a realistic task scenario as they are few (often only one per sense) and not randomly sampled from a corpus. Hence, we decided to create a new diachronic use-sense annotated dataset for the German language.

2.2 Word meaning tasks and models

Currently, several tasks aim to model the meaning of words by utilizing external resources or considering the contexts in which the words appear. These tasks include WSD, WSI, WiC, or LSCD among others.

The WSD task asks to disambiguate a word use, i.e., to select the correct sense for a target word use from all senses of this word listed in some kind of sense inventory (e.g. WordNet) [23]. Gold data for WSD is usually given by use-sense annotated data (see above). In general, this task requires leveraging various linguistic features, such as surrounding words, syntactic structures, and semantic information, to resolve the ambiguity accurately. There are different ways to approach the WSD task including supervised and unsupervised approaches. WSI asks to cluster word uses into clusters corresponding to distinct senses of this word without relying on any predefined sense inventory [5]. It is usually modeled in an unsupervised way.

WiC is a binary classification task that asks to predict whether a word in two uses has the same meaning or not [45]. Gold data for WiC is usually given by use-use annotated data. Several models have been developed to tackle the WiC task, ranging from static word embeddings to more advanced contextualized models. Early approaches using static embeddings faced challenges in accurately determining the similarity or dissimilarity between different contexts due to the limitation of representing multiple word meanings. This deficiency, known as meaning conflation, hindered their ability to capture the nuanced variations in word senses [46]. In contrast, contextualized models [e.g. 20] have demonstrated superior performance by considering each context as a distinct representation, enabling them to capture the subtle differences in word meaning. These models leverage contextual information to generate more fine-grained representations, leading to improved accuracy in the WiC task [45, 47].

The task of detecting words that change their meaning over time is called Lexical Semantic Change Detection [13]. Various models have been developed to address this task, including static and contextualized embeddings. Notably, WiC models have shown remarkable performance in this task, achieving the best results in recent studies [18, 19, 48, 49].

The literature review underscores that despite the extensive research devoted to modeling tasks like WiC, WSI and LSCD, a comprehensive understanding of their interconnections remains elusive. Notably, the absence of diachronic-historic word sense-annotated data presents a significant challenge to conducting a thorough analysis in this domain. To our knowledge, this study represents a pioneering effort to bridge this gap.

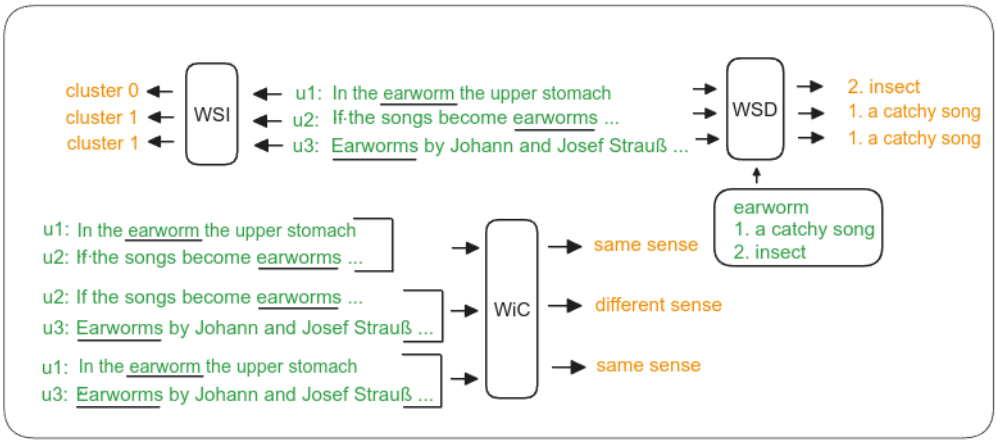


Fig. 1: Visual comparison of tasks involved in LSCD. $u1$, $u2$, and $u3$ represent word use samples. Arrows represent use-sense assignments for WSD, cluster assignments for WSI, and pairwise judgments for WiC.

2.2.1 How Tasks Relate to Each Other

Note that there is a close conceptual connection between WSD, WiC, WSI and the various definitions of LSCD [50] as the process to derive gold labels for LSCD requires at least one of the other tasks to be solved. For instance, Schlechtweg et al. [13] employ a WiC and a WSI step in their annotation process while Basile et al. [40] employ a single WSD step and Schlechtweg et al. [29] employ a single WiC step. All of these approaches have in common that they require some form of sense distinctions: WiC asks to distinguish senses of use pairs, WSI asks to distinguish senses within sets of uses, and WSD asks to assign senses to individual uses. Hence, because of this sense distinction information, they can all be applied to measure sense changes over time. We summarize the relations between the tasks in Figure 1.

Based on the above analysis, we hypothesize that the WSI task can serve as a subtask for solving the LSCD task. By clustering different contexts of a word and employing appropriate metrics, we can quantify the extent to which a word's meaning has shifted from one time period to another, thereby incorporating the dimension of time into our analysis.

3 Datasets

For annotation we chose the existing LSCD dataset DWUG DE [32], which was annotated within the use-use paradigm (see Section 2). This had the advantage that the target words and their uses had already been sampled. We only had to get sense definitions. The dataset has the additional advantage that word uses were randomly

Data	n	N/V/A	|U|	AN	|J|	KRI	STYLE
DWUG DE	50	34/14/2	≤100+≤100	8	1.7	.67	use-use
DWUG DE Sense	24	16/7/1	25+25	3	2.9	.87	use-sense

Table 1: Statistics for the latest version (2.3.0) of DWUG DE and the new DWUG DE Sense dataset. Both datasets contain German word uses from two historical corpora covering 1800–1899 and 1946–1990 respectively. n = no. of target words, N/V/A = no. of nouns/verbs/adjectives, |U| = no. uses per word (t_1+t_2), AN = no. of annotators, |J| = avg. no. judgments per annotation instance, KRI = Krippendorff's α, STYLE = annotation style.

sampled from historical/diachronic corpora.[3] DWUG DE has been annotated in multiple rounds corresponding to different published versions of the data set. We sampled uses for annotation from Version 1.0.0.[4]

3.1 DWUG DE

DWUG DE contains German word uses from 2 time periods annotated with use-use judgments from multiple annotators. The authors sampled pairs of word uses such as (1) and (2) from two historical corpora (1800–1899, 1946–1990) and asked annotators to rate them on a relatedness scale from 1 (unrelated) to 4 (identical).

(1) Im **Ohrwurm** ist der obere Magenmund inwendig mit einigen Zähnen in zwey Reihen besetzt.
*'In the **earwig** the upper stomach mouth is occupied inside with some teeth in two rows.'*

(2) Werden die Lieder **Ohrwürmer**, klingelt auch die Kinokasse.
*'If the songs become **catchy tunes**, the cinema cash register also rings.'*

The annotated pairs were represented as a weighted graph and clustered with Correlation Clustering. All uses sharing a cluster were then interpreted as having the same sense and the semantic change for each word was measured based on these clusterings. Some statistics for the dataset are shown in Table 1.

3.1.1 DWUG DE Sense

We randomly chose 24 target words (out of 50) from the DWUG DE dataset and extracted sense definitions from two historical dictionaries [8, 43].[5] We merged the main sense definitions (no sub-sense definitions) from both dictionaries and included multiple definitions of the same sense by choosing the one that seemed clearer (better

[3] Another positive side effect of relying on an existing dataset is that we have an independent annotation for the same data, allowing for comparisons.
[4] https://zenodo.org/record/5543724
[5] In order to reduce the annotation load, we sub-sampled both target words and their uses.

comprehensible) to one of the authors. Figurative meanings listed in DWDS [43] were treated as separate senses.

We then randomly sampled 50 uses for each target word (25 per time period from at most 100 in the original dataset) and asked three annotators to label each use with a sense definition best describing meaning of the target word in this use.[6] The annotators were asked to assign the label 'andere' ('other') to the use if none of the definitions listed for the target word was suitable. Also the annotators had the option to skip examples, e.g. if they were ambiguous or unclear to them. One annotator is a professional computational linguist, another one holds a degree in linguistics and the third annotator was a current university student with German as a major subject. The annotators had no access to the data before the annotation. In the first round, only the computational linguist annotated the data. This annotator provided additional sense definitions for four words (*abdecken*, *Fuß*, *Manschette*, *Schmiere*) because she deemed the provided definitions insufficient for some uses. These additional definitions were then added to the previous ones and presented indistinguishably to the two other annotators in the second round of annotation.[7]

Consider the following annotation example from the final annotated dataset for the word *Ohrwurm*:

> **use**: *Im* ***Ohrwurm*** *ist der obere Magenmund inwendig mit einigen Zähnen in zwey Reihen besetzt.*, 'In the **earwig** the upper stomach mouth is occupied inside with some teeth in two rows.'
> **sense1**: *Insekt, (von) dem der Volksglaube annimmt, daß es gerne Schläfern ins Ohr kriecht*, 'insect, which is popularly believed to like to crawl into the ear of sleepers'
> **sense2**: *eingängige Melodie*, 'catchy melody'
> **sense3**: *Zuträger, Schmeichler*, 'informer, sweet-talker'
> **sense4**: *andere*, 'other'

This particular example was annotated with 'sense1' by all three annotators. This is different for the following example of *abgebrüht*:

> **use**: *[. . .] das war ein finsterer Herr mit dem harten Blick eines* ***abgebrühten*** *Schellfisches.*, '[. . .] that was a sinister gentleman with the hard look of a **blanched/hard-nosed** haddock.'
> **sense1**: *mit kochendem Wasser übergoßen*, 'doused with boiling water'
> **sense2**: *gefühllos, frech, dickfellig, abgehärtet gegen sittl. Eindrücke etc.*, 'callous, insolent, thick-skinned, hardened to moral impressions, etc.'
> **sense3**: *andere*, 'other'

Two annotators assigned 'sense1' in this example, while one annotator assigned 'sense2'.

Find an overview of the annotated data in Table 1, including a comparison to the original DWUG DE dataset. DWUG DE Sense has roughly half the number of target

[6] The data for some words has multiple repeated uses, e.g. *Ohrwurm*. Similarly, some uses contain a target word several times, and we treat each of these occurrences of the target word as a separate use. Hence, the same sentence may occur various times with the same target word in different places, e.g. for *Abgesang*. We did not remove these instances in order to maintain the data set as realistic as possible.

[7] The annotated data with all derived labels is available at https://doi.org/10.5281/zenodo.8197552.

	A	B	C	full	old	new
A		.84	.89			
B			.89			
full				**.87**	.83	.90

Table 2: Agreement (Krippendorff's α) between annotators on sense definition annotation. Left: Pairwise agreement between annotators on full data. Right: Overall agreement between annotators per time period.

words of DWUG DE and between $\frac{1}{3}$ and $\frac{1}{4}$ the number of annotated of uses per word. However, it has a considerably higher average number of judgments per annotation instance (2.9 vs. 1.7) while employing much less annotators (3 vs. 8).[8] This stems from the lower annotation load per use in the use-sense annotation schema (see Section 2). The higher judgment density brings higher reliability of aggregated data and better possibilities of data cleaning, i.e., removing instances with low agreement (see Section 3.1.1).

Agreement

We present the pairwise agreement of the three annotators in Table 2 as well as the overall agreement (full), where 134 judgments assigning 'other' are ignored. Krippendorff's α is 0.87 for all three annotators and 0.84/0.89 for pairwise agreements, which can be interpreted as quit good agreement. Percentage agreement (ITA) and pairwise Cohen's Kappa [51] yield similar scores with 0.88 and 0.87 for mean pairwise agreement. According to Erk et al. [3], sense annotation studies relying on the WordNet sense inventory show percentage agreement from .67 to .78. Hence, we observe higher agreement than previous synchronic studies. Note, however that agreement also depends on variables such as the granularity of the sense of inventory.[9] The agreement for diachronic use-use annotated data ranges between 0.52 [36] and 0.67 [32].[10] We conclude that our dataset is sufficiently reliable to serve as a gold standard.

We now compare the inter-annotator agreement between the two time periods in order to understand whether the annotation task is harder on the old versus the new data. Please find a comparison of the overall agreement per time period in Table 2 (right). The agreement for the old time period is considerably lower than for the new time period (.83 vs. .90). This indicates that the task of assigning sense definitions is harder for historical than for modern data. This result addresses our initial research

[8] The average number of judgements per instance is lower than 3 because a small number of uses were skipped by some annotators during the annotation.
[9] Note also Section 3 for factors potentially influencing the agreement.
[10] However, these are annotated on 4-point scale which may not be comparable to our binary annotation.

Data	|U|	|S| full	|S| old	|S| new
maj_2	23+23	2.9	2.3	2.5
maj_3	16+18	2.8	1.8	2.4

Table 3: Statistics of cleaned datasets. $|U|$ = average number of uses per time period, $|S|$ = average number of senses (in all cases the median number of senses is 2).

question (outlined in Section 1) by showing that there is a tendency for human annotators to show higher agreement on modern data. This is in line with expectations, as annotators, despite their expertise, are inherently more familiar with modern data.

It also suggests that computational models may struggle similarly more with the historical than with the modern data. Note though that by cleaning the data (as described in Section 3.1.1) this difference between historical and modern data should be considerably decreased.

Cleaning and label derivation

After the annotation process we created two different cleaning conditions: (i) We removed all instances for which all annotators assigned different labels. We also removed uses with at least two missing annotations, leaving 1117 uses (from 1200). For each use, we then chose the sense definition assigned by the majority of annotators (at least two) as gold label. We call this cleaned version of the data 'maj_2' (majority label with agreement of at least two annotators). (ii) We removed all instances where not all 3 annotators annotated the same label. We also removed uses with missing annotations, leaving 826 uses. To each use, we assigned the label chosen by the majority of annotators (in this case by all three). Hence, we call this version of the dataset 'maj_3' (majority label with agreement of all three annotators).

We refer to the subset of annotated data corresponding to the old/new time period as 'DWUG DE Sense Old/New' respectively. Sense clusters for WSI are given by the aggregated gold sense labels from above. From the extracted clusters we compute sense frequency distributions and infer binary and graded change labels as described in Schlechtweg et al. [13]. The binary change score measures whether a sense was gained or lost over time, or not.[11] The graded change score corresponds to the Jensen-Shannon distance between sense probability distributions. As an example, consider Figure 2. On the left, we see the sense frequency distribution for the full (both time periods) data for the word *Ohrwurm* which was aggregated and cleaned with the maj_3 condition. *Ohrwurm* has two senses with rather balanced frequencies in the full data (19 vs. 16 uses). However, in the old and new portion only one of these senses exists respectively. Hence, the word lost a sense and gained another sense (binary change). No sense exists across the two time periods, i.e., the word completely changed its

[11] Only senses reaching a certain occurrence frequency are considered. Find the code used to aggregate and clean the data, and to derive proximity and change labels at https://github.com/Garrafao/WUGs.

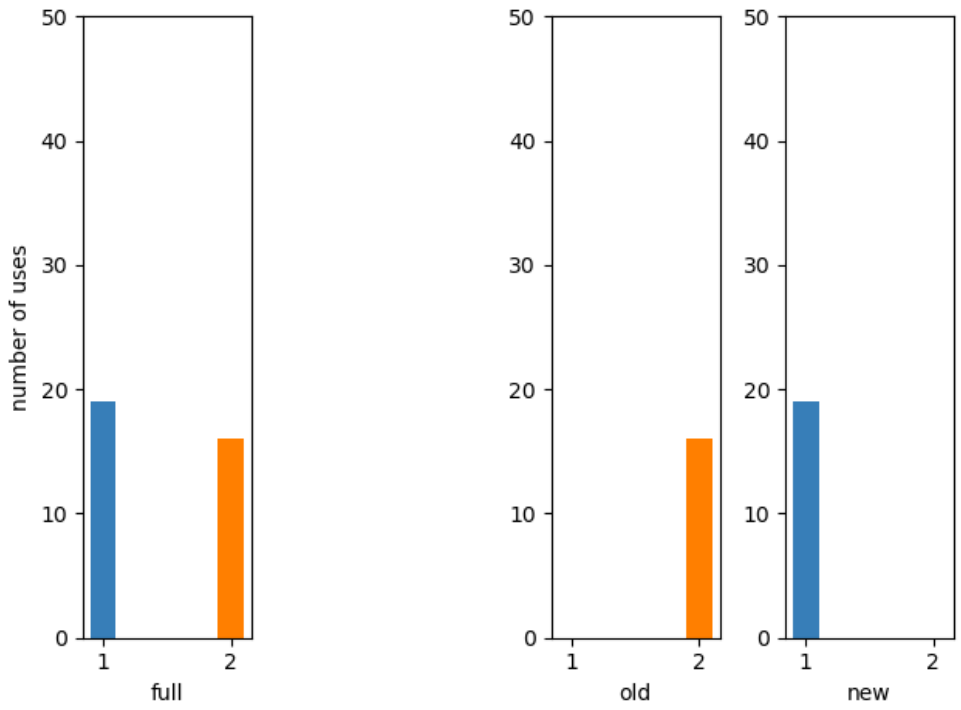


Fig. 2: Sense frequency distribution for *Ohrwurm* from cleaned dataset (maj_3).

meaning and has a Jensen-Shannon distance of 1.0 (see also Figure 5). *Manschette* in Figure 3, instead, has a skewed sense frequency distribution in the full data with three senses and the most frequent sense dominating strongly in frequency. In the old portion of the data, only the dominating sense exists while in the new portion the two less frequent senses occur with comparably low frequencies. Hence, these two senses are gained over time. The Jensen-Shannon distance is medium-high with 0.46 reflecting the low frequency of the gained senses.

For each of the two cleaning conditions (maj_2, maj_3), we also infer binary WiC (or semantic proximity) labels for all use pairs per word in a procedure similar to the one described in Pilehvar and Camacho-Collados [45], i.e., pairs of uses with the same sense definition as majority label receive label '1' while those pairs with different sense definitions receive label '0'. This means that use pairs from our dataset also have WiC annotations and can be used for evaluating WiC models.

Please find some statistics of the cleaned version of the dataset in Table 3, respectively. The cleaning leads to a slight drop of the average number of uses per time period for maj_2, and a larger drop for maj_3. The average number of senses is 2.9 and

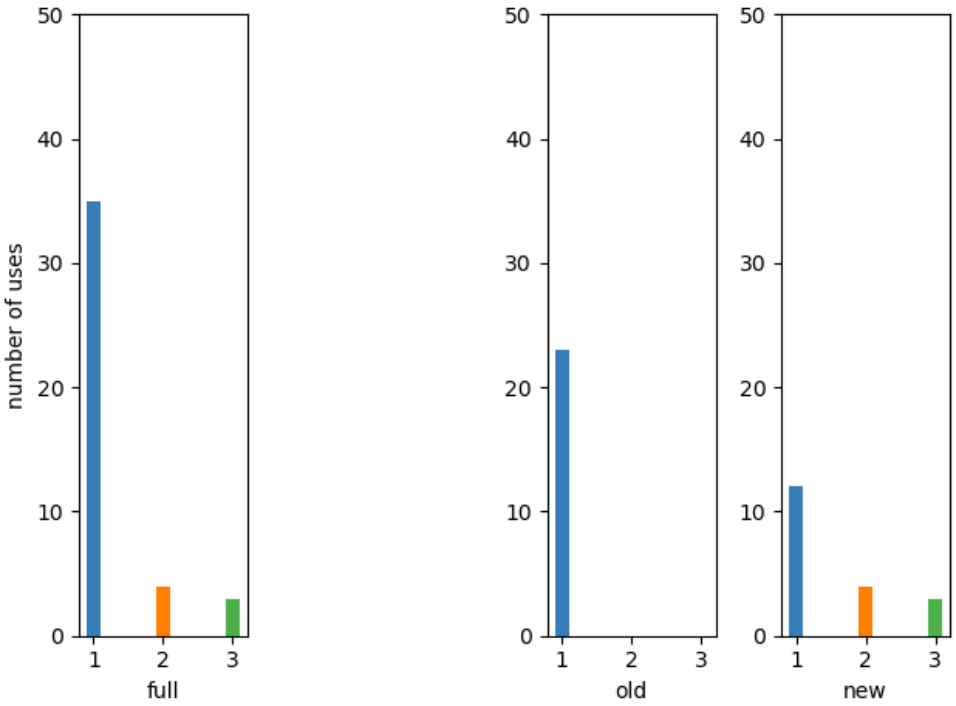


Fig. 3: Sense frequency distribution for *Manschette* from cleaned dataset (maj_3).

2.8 for different cleaning conditions on the full data. On all data splits (full/old/new), the average sense number is lower for maj_3 as a result of the stronger cleaning. For the old portion, the difference between the portions is most pronounced (2.3 vs. 1.8) due to higher disagreement between the annotators. Figure 4 shows the number of senses per word for maj_3 on the full data. Most words have two senses across time periods and there are individual exceptions with 1, 6 and 7 senses respectively (*artikulieren*, *Schmiere*, *Fuß*). Figure 5 shows the Jensen-Shannon distance between cluster probability distributions for each word for maj_3, i.e., the graded change values inferred on the annotated and cleaned data. As we see, they are well-distributed across the possible values between 0.0 and 1.0. Some words have complete change (*Ohrwurm*, *Seminar*) while others have no change at all (*artikulieren*).[12]

[12]Note that this is after cleaning. So, with different (less strict) cleaning conditions the change values may be different.

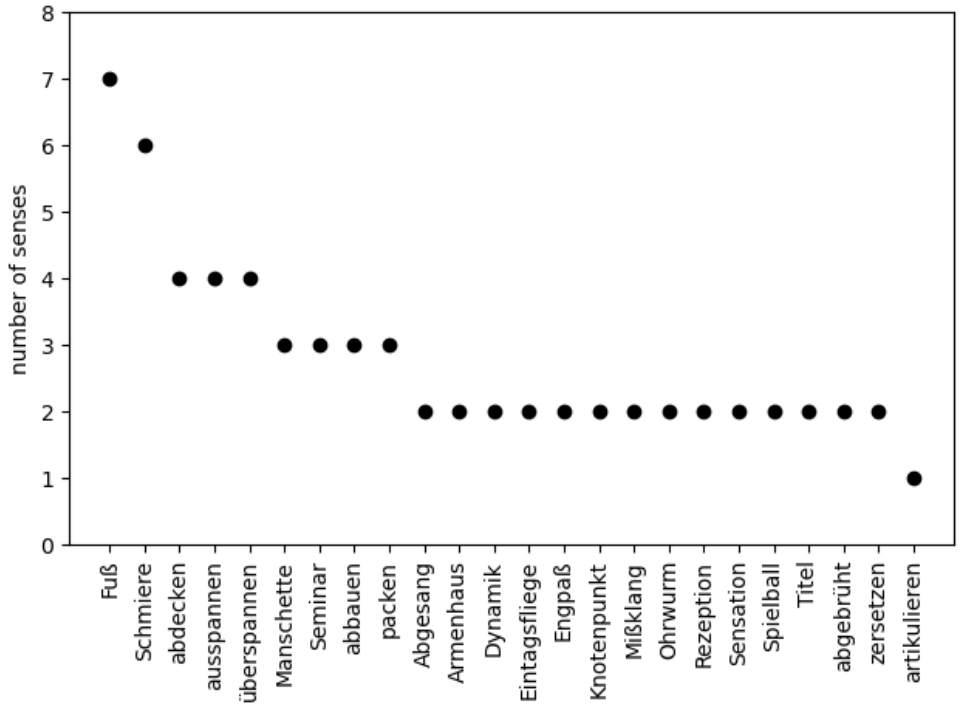


Fig. 4: Number of senses per word for cleaned dataset (maj_3).

4 Tasks

The aim of our study is to establish a coherent relationship among word meaning models, particularly when applied to diachronic-historical word sense annotated data, as discussed in Section 1. With the creation of a dataset tailored to these specifications, this section provides a more precise definition of the two tasks central to our investigation: WSI and LSCD.

4.1 WSI

The WSI task is an unsupervised task that considers grouping uses corresponding to different senses of a given target word into clusters or groups. These clusters are interpreted as individual senses considering the context in which the given target word occurs [52, 53], and do not have to correspond to a predefined set of senses for evaluation. It is closely related to WSD, but with a crucial difference: WSD requires a predefined sense inventory for each word, making it a supervised task. We chose to focus on WSI rather than WSD in our experiments due to the limited size of our

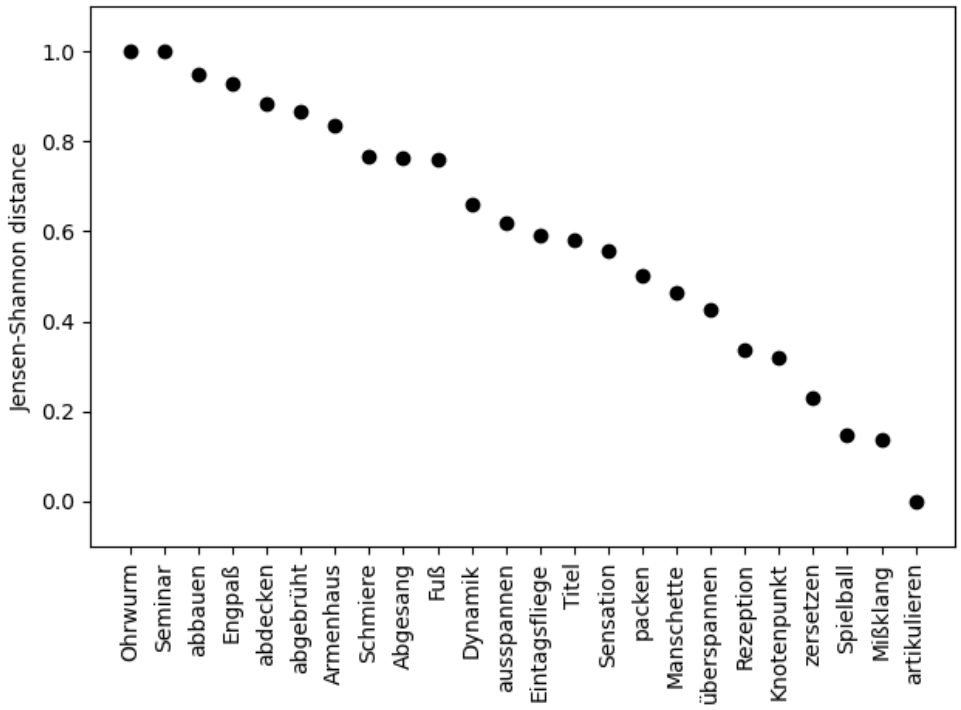
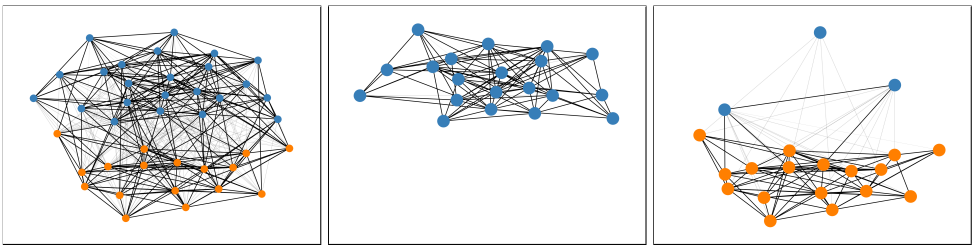


Fig. 5: Jensen-Shannon distance between sense probability distributions for cleaned dataset (maj_3).

dataset. With only a few target words associated with a small number of contexts, the dataset is considered insufficient for effectively addressing WSD. However, the high quality annotations shown in Table 2 allow us to derive cluster labels using the target word senses, making our data suitable for the task of WSI.

Our dataset differs from more standard WSI and WSD datasets [54] by its inclusion of sense annotations of uses coming from two different time periods. As mentioned in Section 1, this unique feature raises important questions about the relationship between sense detection and semantic temporal drifts. Specifically, we can investigate whether WSI systems yield improved results during certain time periods compared to others. Moreover, utilizing the WSI task on diachronic word-sense annotated data offers a more detailed understanding of the system's capability to detect semantic changes.

We use the Adjusted Rand Index (ARI) [55] to measure systems' performance on this task. ARI is essentially a measure of the similarity between two clusterings (partitionings) of a dataset. It is used to evaluate the performance of clustering algorithms by



$G,\ D = (23, 17)$ $G_1,\ D_1 = (20, 0)$ $G_2,\ D_1 = (3, 17)$

Fig. 6: Word Usage Graph from Zamora-Reina et al. [19], for the word *servidor* (left), subgraphs for old corpus G_1 (middle) and for modern corpus G_2 (right). The colors correspond to the clusters. **black**/gray lines indicate **high**/low edge weights.

comparing the agreement between the predicted clusters and the true clusters (ground truth) of the data. A noteworthy property of this metric is that it does not requiere predicted and gold clusters to be compatible in terms of the number of categories as it operates at the level example pairs in its calculation.

4.2 LSCD

The LSCD task has been defined differently in the literature. For this work, we adopt the widely accepted graded view of the task presented in [13, 19], which is described below:

> Given a set of target words and a set of uses U_1 and U_2 for each of them, the task is to rank the target words according to their degree of LSC between U_1 to U_2.

The gold standard rankings are computed using the Jensen Shannon distance of the sense frequency distribution for each word between the two time periods, thereby generating a list of words sorted by the degree of semantic change [13, 19]. Finally, the evaluation metric used is the Spearman correlation [56] between the gold and predicted word rankings. Our dataset is well-suited for addressing this task, as it provides the necessary information for computing the sense frequency distribution for each word in both time periods and subsequently constructing the graded change rankings. This allows to evaluate LSCD systems based on their correct sense assignments for two different time periods, enabling an assessment of their ability to compute sense distribution changes. We argue that this approach provides a more comprehensive understanding of how well these systems capture and comprehend semantic shifts over time. A central research question of this study is to investigate the relationships between these two tasks in terms of how systems perform across them.

5 Models

This section provides a detailed description of the model architectures that we employ in our experiments to solve the tasks defined above. We present each component part of the architecture and provide a comprehensive explanation of its role in the overall system. By doing so, we aim to provide a clear understanding of how the system operates and how it was designed to address the specific research questions of interest.

Our architecture is based on state-of-the-art models applying components used in one of the recent LSCD shared tasks [13, 19, 48]: First, we use a WiC model [18], as described in Section 5.1, to generate predictions for pairs of uses that assess the similarity of a target word based on its contextual information. Second, the WiC model predictions are used to construct a weighted graph, called *Word Usage Graph* (WUG) [32], where nodes represent uses and edges represent the probability that two uses are similar. Finally, clustering algorithms are applied to the WUG, producing clusters that correspond to the different senses of the target word.[13] Overall our pipeline is based on three main steps (see Figure 7):

1. apply WiC model to score all pairs of word uses of each target word separately,
2. build the WUG, i.e., a graph with nodes corresponding to different uses of the same target word and edges weighted with WiC scores for the corresponding pairs of uses,
3. apply clustering algorithms to word uses to potentially establish meaningful clusters representing word senses.

From the full cluster graph, we can construct two time-specific subgraphs with two time-specific frequency cluster distributions, from which we calculate graded change predictions with Jensen Shannon distance (see Section 3). Find an example of a clustered WUG in Figure 6. It shows the full graph G on the left and the time-specific graphs G_1 (old) and G_2 (new) with their respective cluster frequency distribution. We describe the three parts of our pipeline with more detail in the subsections below.[14]

5.1 Word-in-Context

The WiC task is to determine if two occurrences of the same ambiguous target word in two different sentences or text fragments have the same or different senses [45]. To solve this task we employ several *DeepMistake* models [18, 57], which had previously shown SOTA or near-SOTA results in two shared tasks [19, 48]. These models were trained as probabilistic binary classifiers predicting the probability that two given occurrences of some target word have the same sense. We use this probability as a measure of similarity between word uses, with higher probability corresponding to

[13] We also tried experimenting with substitution-based state-of-the-art models, find more details in Appendix A.

[14] The architecture of our pipeline implies that there are considerable differences between the derivation of the gold clusters (Section 3.1.1) and model clusters: Gold clusters are derived from use-sense judgments not requiring clustering as we filter out all ambiguous judgments. Model clusters are derived from (pairwise) use-use similarity predictions from a WiC model with an additional clustering algorithm. LSCD is then indeed measured in the same way from gold and model clusters. Hence, even if the WiC model were the ideal approximator of human (WiC/use-use) judgements, then model WSI and LSCD results are not bound to be perfect because the ground truth was not constructed with a clustering of use-use judgments, but from use-sense judgments. Of course, there is a certain correspondence between human use-use and use-sense judgments, but this correspondence can be ambiguous.

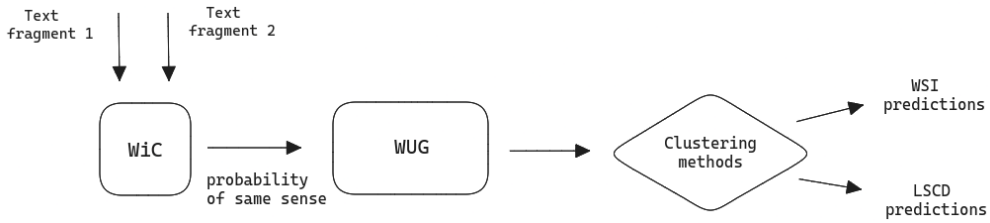


Fig. 7: Overview of the architecture used to generate predictions for the WSI and LSCD tasks.

more similar uses. All models we employ have the same architecture, which was shown to perform best in the previous studies. The difference between the models lies in the data they were trained on. Since none of the DeepMistake models was trained on any labled WiC data in German, we take the models that had previously shown the best results in the LSCD shared tasks in Russian and Spanish, which were trained on a mixture of diachronic and synchronic WiC data in these languages. For comparison, we additionally employ the models trained on a synchronic WiC dataset in 5 languages, and also on subsets in English or in Russian only. In total, 6 WiC models were involved in the process of hyperparameter selection.

5.1.1 The architecture of DeepMistake

Given a pair of uses of some target word, these uses are concatenated and encoded by the XLM-R large backbone [58]. Specifically, the backbone gets the input in the following format: `<s>usage1</s>usage2</s>`.[15] For each of two occurrences of the target word the outputs from the last Transformer layer at the positions of subwords of the target word are averaged (i.e. mean pooling over subwords of the target word is done). Thus, we get two contextualized embeddings for two occurrences of the target word.

Then two contextualized embeddings are combined and passed to the classification head. Following [18], we concatenate the L1-distance and the dot product between the normalized embeddings: $(\|\bar{x} - \bar{y}\|_1, \langle \bar{x}, \bar{y} \rangle)$. The classification head consists of a batch normalization and a linear layer followed by the softmax over two classes. Models were trained on several WiC datasets using the CE loss. Both the XLM-R backbone and the classification head were fine-tuned.

Taking into account the symmetric nature of the same-sense relation, i.e. whether two occurrences of some target word have the same sense or different senses should not depend on the order in which these occurrences appear, DeepMistake employs both training time and test time augmentation. During training, for each

[15] When an input is longer than $L = 500$ subwords, we shorten it to be exactly this length. Specifically, we shorten left and right contexts of each usage that are longer than $L/4$ subwords proportionally to their extra length and leave those that are shorter than $L/4$ intact.

Data source	Languages	size	#targets	Avg. len.
MCL-WiC	en-en	8008	3728	48
	ru-ru	708	352	41
	fr-fr	708	352	46
	ar-ar	708	354	45
	zh-zh	708	342	-
	en-nen*	32	16	51
RuSemShift	ru-ru	3898	70	51
Spanish DWUG	es-es	5443	15	167
Spanish XL-WSD	es-es	8260	310	98

Table 4: Training data of the employed DeepMistake models. Each example is a pair of uses of the same word in the same language, except for en-nen* which are cross-lingual examples with one usage in English and another in French, Russian, Arabic or Chinese.

`<s>usage1</s>usage2</s>` a symmetric example `<s>usage2</s>usage1</s>` was automatically generated, thus increasing the number of training examples. During inference for each edge in a WUG two (ordered) pairs of uses are generated and scored by DeepMistake, then the scores are averaged to get the final edge weight.

5.1.2 WiC training and model variants

In this work we experimented with several DeepMistake models differing in how they were trained and which data they were trained on. The training examples were taken from several source datasets which are summarized in table 4.

1. **MCL-WiC** is a synchronic Multilingual and Cross-Lingual Word-in-Context dataset [59]. For training the authors of DeepMistake employed the original training set in English, 70%[16] of each development set in French, Russian, Arabic, and Chinese, and also all the examples from the trial sets including few cross-lingual examples.
2. **RuSemShift** is a diachronic dataset proposed in [36] and consisting of pairs of uses in Russian from pre-Soviet, Soviet and post-Soviet epochs. The original labels are values from 1 to 4. For training DeepMistake all pairs with labels 3 or larger were treated as positive examples and the rest as negative ones. Compared to the Russian part of MCL-WiC, RuSemShift contains 5x more examples but 5x less different target words.
3. **Spanish DWUG** is a diachronic dataset in Spanish from [19] containing pairwise human annotations from 1 to 4 similarly to RuSemShift. For training DeepMistake only pairs with the gold scores of 1 and 4 were employed as negative and positive examples correspondingly, all other pairs were filtered out. This kind of filtering was shown to be beneficial for the model performance [18].

[16] The rest 30% were used as validation data for early stopping.

4. **Spanish XL-WSD** is the Spanish part of the synchronic XL-WSD Word Sense Disambiguation dataset [60], which was converted to the WiC format. Specifically, the Spanish development and test subsets of this WSD dataset were taken, and all pairs of examples sharing the same target word lemma were generated. Those pairs having identical sense labels for both usages of the target word were labeled as positive, and the rest as negative. Compared to the Spanish DWUG, this dataset suggests much higher diversity of target words.

We experimented with the following DeepMistake models.

1. **MCL→RU** (or "mean+dist_l1ndotn-hs0 on $\text{MCL}_{CE}^{nen-acc}$ →$\text{RSS}_{CE}^{dev2-sentSpear}$" in the notation of [57]) is the DeepMistake model that has shown the best results on the Russian RuShiftEval-2021 dataset[17]. It was fine-tuned in two stages, first on the multilingual MCL-WiC dataset and then on the diachronic RuSemShift dataset in Russian.
2. **MCL→ES** (a.k.a. "MCL→DWUG_es$_{ALL}^{bin2}$+XL-WSD" in [18]) is the DeepMistake model that has achieved the best Spearman's correlation with the gold COMPARE scores among all participants of the Spanish LSCDiscovery-2022 shared task. The first stage of fine-tuning is identical to the previous model, but the second stage was performed on both synchronic and diachronic data from Spanish DWUG and Spanish XL-WSD.
3. **ALL** (a.k.a. "MCL+RSS+DWUG_es$_{ALL}^{bin2}$+XL-WSD" in [18]) is the model that has achieved the best Spearman's correlation with the gold JSD scores among all DeepMistake models in LSCDiscovery-2022. It was fine-tuned on all training data listed in table 4 in a single stage.
4. **MCL** (a.k.a. "MCL-WiC(CE)" in [57]) is fine-tuned in one stage on all the MCL-WiC data in 5 languages.
5. **enMCL** (a.k.a. "MCL-WiC_train-en-en(CE)" in [57]) is fine-tuned in one stage on the English part of MCL-WiC only.
6. **ruMCL** (a.k.a. "MCL-WiC_ru-ru(CE)" in [57]) is fine-tuned in one stage on the examples from MCL-WiC in Russian only. Additional 1000 examples from the test set in Russian were added, resulting in 1708 examples in total.

5.2 Building the Word Usage Graph

With the predictions of the WiC model, a complete graph is constructed in which the nodes represent uses and the weighted edges denote the similarity between pairs of uses. The resulting graph serves as a structural representation of the dataset that allows clustering algorithms to identify different senses of a target word based on the weighted edges. We explore different mechanisms for constructing the graph, which can be expressed through the following hyperparameters:

[17] This is an improved version of the best model from [57], it has achieved the average Spearman's correlation with the gold COMPARE scores across all pairs of time periods equal to 0.85, thus, outperforming the winner of the RuShiftEval-2021 competition. See https://github.com/Daniil153/DeepMistake/ for all metrics of this model.

Parameter	Values
`binarize`	[True, False]
`quantile`	[1,...,10]
`WiC model`	[enMCL, ruMCL, MCL→RU MCL, MCL→ES, ALL]
`fill_diagonal`	[True, False]
`use_disconnected_edges`	[True, False]

Table 5: Hyperparameters to build a WUG.

- `threshold`: All edges with weights less than this hyperparameter are dropped. During grid search, it acquires the values of the 10%, 20%, ..., and 90% percentile of all edge weights, along with a default value of 0.5.
- `binarize`: When set to true, all edges are assigned a weight of 1 (subsequent to dropping the edges with weight less than `threshold`).
- `fill_diagonal`: This parameter controls if the main diagonal of the adjacency matrix is filled with 1.0 or 0.0 as value.
- `use_disconnected_edges`: When set to true, edges with a weight less than the cutoff threshold will be kept as part of the network with a weight of 0.0. For Chinese Whispers and Spectral Clustering (see Section 5.3), this value will default to True, for WSBM and Correlation Clustering we will only keep them in case of binarization.[18]

We additionally scale the weights of each graph with a MinMaxScaler fitted on the full set of similarity predictions of the respective WiC model because scores are tightly center around 0.5. Table 5 shows a summary of the hyperparameters needed to build a WUG.

5.3 Graph Clustering

In the final step of our pipeline, we apply graph clustering methods to our WUG to obtain word senses in an unsupervised fashion. Below, we describe all the graph clustering methods that we consider in this study. Additionally, Table 6 provides a comprehensive overview of the clustering methods along with the parameter ranges applicable to each model.

5.3.1 Chinese Whispers

Chinese Whispers (CW) is an efficient, randomized clustering algorithm with a time complexity linear with respect to the number of edges [61]. The algorithm first assigns all nodes to different clusters. Then the nodes are processed in randomized order for

[18] 0-weighted edges are removed for WSBM and Correlation Clustering (without binarization) because otherwise wrongly influence the clustering algorithm.

Clustering method	Parameter	Values
Chinese Whispers	`weighting`	[lin, top, log]
Correlation Clustering	`threshold_cc`	0.5
WSBM	`distribution`	[binomial, poisson, geometric, normal, exponential]
Spectral Clustering	n_c	[Silhouette, Calinski-Harabasz, Eigengap]

Table 6: Clustering method hyperparameters.

a small number of iterations (we set this hyperparameter to 20) and are assigned to the strongest cluster in the local neighborhood, i.e., the cluster whose sum of edge weights to the current node is maximal. The calculation of edge weights is controlled by the `weighting` hyperparameter, which takes three different values:

1. lin: This calculates the weight of an edge between two nodes in a graph using linear weighting, which is the edge weight divided by the degree of the destination node.
2. log: This computes the weight of an edge between two nodes in a graph using logarithm weighting, which is the edge weight divided by the logarithm of the degree of the destination node.
3. top: This keeps edge weight as is.

We use the implementation provided by Ustalov et al. [62].

5.3.2 Correlation Clustering

We use a variation of Correlation Clustering (CC) [63], a graph clustering technique which minimizes the sum of cluster disagreements, i.e., the sum of negative edge weights within a cluster and the positive edge weights across clusters [13]. This method has, to our knowledge, not previously been employed for the WSI task. However, it has been used extensively in the LSCD context to cluster human annotations [19, 32, 37, 38, 64–66], but also to cluster model predictions [18]. Correlation Clustering has hyperparameters for splitting edge weights into positive and negative (`threshold_cc`), for the maximum number of senses (set to 10), and for maximum attempts and maximum iterations for simulated annealing (set to 2000 and 50000 respectively). `threshold_cc` is set to 0.5. It is worth noting that `threshold_cc` is normalized with the same scaler as applied to all edge weights (see above).

5.3.3 Weighted Stochastic Block Model

We use a Bayesian formulation of the Weighted Stochastic Block Model (WSBM), a generative model for random graphs popular in biology, physics and social sciences [67, 68]. The model has been first applied on WUGs by Schlechtweg et al. [33] and subsequently by Kotchourko [69] and Tunc [70]. The basic assumption of the WSBM is that nodes belong to latent blocks (clusters), and that nodes in the same block are stochastically equivalent (i.e., they have edges drawn from the same distribution). Fitting the model is equivalent to determining the optimal latent block structure providing a clustering of word uses.[19] We use the WSBM version described in Schlechtweg et al. [33] marginalizing out edge probabilities.

The WSBM model supports several distributions from which edge weights are drawn through the `distribution` parameter which can take one of the following values: exponential, normal, poisson, binomial, and geometric.[20] Exponential and normal distribution are only applicable to real-valued edge weights while poisson, binomial, geometric distribution are only applicable to discrete edge weights. Hence, we apply the former two distributions only to non-binarized graphs while we apply the latter three distributions only to binarized graphs.

5.3.4 Spectral Clustering

Spectral Clustering (SC) is a class of algorithms that apply clustering to a low-dimension projection of the affinity matrix of the graph [71]. We use the scikit-learn[21] implementation with default hyperparameters. We apply the K-means algorithm to find clusters in the reduced-dimensional space. This algorithm requires the number of clusters as input parameter. Next, we explain the methods employed for selecting the number of clusters.

Silhouette Score

This is defined as the mean silhouette coefficient of all nodes. The coefficient is the difference of the mean nearest-cluster distance and the mean intra-cluster distance, divided by the maximum of the two [72].

Calinski-Harabasz Score

The score is defined as ratio of the sums of between-cluster dispersion and of within-cluster dispersion [73].

Eigengap heuristic

This technique selects the number of clusters k as the value which maximizes $\lambda_{k+1} - \lambda_k$, where $\lambda_1, \lambda_2, \cdots, \lambda_k$ are the eigenvalues of the affinity matrix Laplacian [71].

In the case of Silhouette and Calinski-Harabasz Score, the clustering with the highest score was selected, as these metrics are designed to favor more representative

[19] https://graph-tool.skewed.de/static/doc/autosummary/graph_tool.inference.BlockState.html
[20] https://graph-tool.skewed.de/static/doc/demos/inference/inference.html
[21] https://scikit-learn.org

Tasks	**WSI**		**LSCD**	
metrics	ARI	SPR	ARI	SPR
Chinese Whispers	(6) .622 ± .130	(6) .768 ± .241	(5) .602 ± .078	(4) .768 ± .241
Correlation Clustering	(5) .622 ± .132	(2) .873 ± .110	(3) .666 ± .096	(5) .715 ± .207
WSBM	(3) .710 ± .068	(1) .911 ± .156	(4) .613 ± .158	(2) .828 ± .169
Spectral Clustering/s	(1) **.771** ± .090	(2) .873 ± .141	(1) .715 ± .067	(1) **.835** ± .130
Spectral Clustering/c	(2) .755 ± .087	(4) .834 ± .077	(2) .702 ± .087	(3) .774 ± .111
Spectral Clustering/e	(4) .633 ± .144	(5) .807 ± .202	(6) .426 ± .208	(6) .596 ± .213

Table 7: Comparison of clustering methods: rank and performance metrics. The table presents the rank (in parentheses) for the cross-validation experiment and corresponding performance metrics (ARI and SPR) for each clustering method. Lower ranks indicate better performance, and the values are reported as means with standard deviations.

clusterings. Additionally, in cases where multiple clusterings had the same score, we selected the one with the fewest clusters.[22] We refer to Spectral Clustering with the respective method to choose the number of clusters by ".../s" for Silhouette, ".../c" for Calinski-Harbasz, and ".../e" for Eigengap.

6 Experiments

In this section, we describe the experiments conducted to evaluate the performance of our model architecture (see Section 5) utilizing the annotated data described in Section 3, which spans different time periods and encompasses annotations for 24 words and 826 uses. We design a series of experiments to measure the generalization capabilities of different systems for the WSI and LSCD tasks. To achieve this, we employ a cross-validation approach with hyperparameter optimization [74].

Our set of 24 words is partitioned into 5 folds. Subsequently, within each cross-validation iteration, an exhaustive hyperparameter grid search is conducted. This encompasses hyperparameters for the WiC model, WUG, and those specific to the targeted clustering method. The optimal configuration identified in the training folds is then evaluated in the testing fold. This process yields 5 performance scores (ARI for WSI and Spearman for LSCD) for each clustering method, and we calculate the mean and standard deviation from these scores. We separately report results using WSI and LSCD performance as optimization criterion. This experimental design is motivated by the objective to unveil the correlation between the WSI and LSCD tasks. Specifically, it investigates whether the WSI task can effectively function as a proxy for the LSCD task, and vice versa.

Table 7 provides a summary of the results obtained from our 5-fold cross-validation experiments. It displays the average test score across folds along with the standard

[22] We use an adjacency matrix representing sentence similarity, rather than a feature matrix, to calculate the Silhouette Score and Calinski-Harabasz Score.

Methods	old ARI	new ARI
Chinese Whispers	.640 ± .170	.578 ± .108
Correlation Clustering	.495 ± .060	.596 ± .219
WSBM	.495 ± .159	.640 ± .070
Spectral Clustering/s	.679 ± .166	.765 ± .040
Spectral Clustering/c	.699 ± .105	.721 ± .151
Spectral Clustering/e	.659 ± .119	.663 ± .104

Table 8: Comparison of the cross-validation performance when testing only on the old or the new portion of the dataset respectively, across all methods.

deviation. The WiC models used as hyperparameters may vary between folds. Tables 9 and 10 detail which WiC model was selected for each fold. Additionally, the reported standard deviation quantifies how the metrics vary from one fold to another, depending on both the hyperparameters selected in each fold and the test subset corresponding to that fold. For Spectral Clustering, we present the results for each method to choose the number of clusters separately. The table presents two distinct approaches to the hyperparameter selection: optimization for WSI (left) and for LSCD (right). For both approaches, we report also the performance on the other task which we do not directly optimize for in order to understand how well the performance translates to the other task. Among the methods evaluated, Spectral Clustering stands out as the most effective, particularly when utilizing the Silhouette score to select the number of clusters. This configuration consistently obtains the highest ARI score considering the WSI task and the highest SPR in the LSCD task.

As we can see in Table 7, WSBM is the most effective model when translating parameters from WSI to LSCD. Conversely, when applying the optimized parameters from the LSCD task to address the WSI task, the Spectral Clustering method, specifically using the Silhouette validation method, demonstrates superior performance, yielding the most favorable results. An additional noteworthy observation is that, with the exception of Chinese Whispers, all methods consistently yield superior results when utilizing parameters optimized for the WSI task to address the LSCD task, compared to using parameters optimized for the LSCD task. We conjecture that this is because the model selection criterion for LSCD overfits more easily because of the notoriously small number of annotated examples (words) compared to the WSI task (uses). This problem affects all existing LSCD datasets [e.g. 19, 32, 38, 75]. Hence, we consider this finding relevant for the LSCD community in general. It suggests that optimization for WSI should be preferred over optimization for LSCD in small data scenarios, even if LSCD is the task of interest.

Table 8 presents the cross-validation results for the old and new dataset partitions (see Section 3). The WiC models used as hyperparameters may vary between folds. Additionally, the reported standard deviation quantifies how the metrics vary from

one fold to another, depending on both the hyperparameters selected in each fold and the test subset corresponding to that fold. With the exception of Chinese Whispers, all other methods exhibit a higher ARI for the new partition compared to the old one. These findings align with the dataset annotations, where inter-annotator agreement is higher for the new portion of the dataset (see Table 2). They also indicate that our cleaning procedure described in Section 3, did not completely remove the additional difficulties introduced by historical language data.

6.1 Results per fold

Tables 9 and 10 display the results of our experiments for each clustering method across individual test partitions as part of the 5-fold cross-validation. Both tables share common columns, including "fill-diag" representing the fill diagonal parameter, "ude" indicating the use of disconnected edges, and "mod-hyp" representing model hyperparameters (see Section 5). Notably, the "test-SPR" column in Table 9 shows the best SPR results considering optimized parameters from the WSI task, whereas the "test-ARI" column in the Table 10 displays the optimal results for the LSCD task, considering parameters optimized in the LSCD task.

In Table 9, we observe that for Spectral Clustering with Silhouette, the selected model is the same for four out of five folds. This indicates that the method remains stable, consistently performing well under the same set of parameters; in contrast, in Table 10 the method shows more variation. We observe a similar picture for WSBM showing stability when optimized for WSI, but less for LSCD.

We observe for both tasks that setting the parameter "ude" (`use_disconnected_edges`) for Correlation Clustering and WSBM to True consistently yields superior results. Regarding the `distribution` parameter ("mod-hyp") of the WSBM, it becomes evident that when set to "poisson" it obtains optimal performance across all five folds for WSI. For LSCD, the geometric distribution dominates in four out of five folds. Furthermore, we observe that for the WiC path parameter, the value "ALL" dominates across all methods for WSI. This pattern, however, is not consistently observed when the models are optimized for the LSCD task.

7 Conclusions and Future Work

We presented a human-annotated, diachronic-historic use-sense dataset which can be used for WSI, WSD, WiC, and LSCD. We found that human annotators show considerably lower agreement on the historical portion of our data than on the modern one (RQ1), which is also reflected in the respective model performance on these portions (RQ2). We conducted experiments testing the performance of various graph clustering models on the dataset with WSI and LSCD. The main finding is the dominance of WSI-based model selection for LSCD. Importantly, attempts at reverse fine-tuning did not yield comparable effectiveness. This means that WSI optimization can indeed yield a model performing optimal on LSCD while the opposite is not true on our data (RQ3). However, we hypothesize that this effect is related to data size and may vanish with larger data sets.

quantile	WiC-path	binarize	ude	mod-hyp	fill-diag	dev-ARI	test-ARI	test-SPR
Method: Chinese Whispers								
6	ALL	True	True	w: lin	True	0.745	0.683	0.975
6	ALL	True	True	w: log	True	0.667	0.809	0.900
6	MCL→RU	True	True	w: log	True	0.714	0.637	0.400
5	MCL→RU	True	True	w: log	True	0.778	0.403	0.564
6	ALL	False	True	w: lin	False	0.670	0.580	1.000
Method: Correlation Clustering								
3	ALL	False	False	-	True	0.757	0.623	0.975
1	ALL	False	False	-	True	0.735	0.746	0.990
3	ALL	False	False	-	True	0.765	0.604	0.900
4	MCL→RU	True	True	-	True	0.757	0.387	0.700
5	enMCL	True	True	-	True	0.729	0.749	0.800
Method: WSBM								
0	ALL	True	True	poisson	False	0.718	0.679	0.975
0	ALL	True	True	poisson	False	0.691	0.781	0.999
0	ALL	True	True	poisson	True	0.727	0.679	0.600
5	ALL	True	True	poisson	True	0.737	0.615	0.999
0	ALL	True	True	poisson	False	0.705	0.794	1.000
Method: Spectral Clustering/s								
4	ALL	False	True	-	True	0.838	0.753	0.975
4	ALL	False	True	-	True	0.818	0.818	0.999
4	ALL	False	True	-	True	0.839	0.754	0.700
3	enMCL	False	True	-	False	0.811	0.629	0.700
4	ALL	False	True	-	True	0.811	0.902	1.000
Method: Spectral Clustering/c								
3	enMCL	False	True	-	False	0.788	0.838	0.872
4	ALL	True	True	-	False	0.794	0.748	0.900
3	enMCL	False	True	-	True	0.818	0.598	0.900
3	enMCL	False	True	-	False	0.811	0.753	0.700
3	enMCL	False	True	-	False	0.791	0.837	0.800
Method: Spectral Clustering/e								
5	ALL	True	True	-	False	0.680	0.652	0.975
0	ALL	True	True	-	True	0.629	0.807	0.900
0	ALL	True	True	-	True	0.643	0.738	0.700
0	ALL	False	True	-	True	0.711	0.390	0.462
5	ALL	False	True	-	True	0.682	0.580	1.000

Table 9: Detailed hyperparameters employed in constructing the WUG and the models during cross-validation, outlined for each fold and model within the context of the WSI task. Each row corresponding to a method signifies a distinct fold.

Furthermore, using Spectral Clustering with Silhouette score to choose the number of clusters yielded top results on both tasks showing surprising stability on WSI.[23]

We hope that this study pave the way for future research in the LSCD field. Likewise, our future work will focus on refining and extending the current findings, aiming for a more comprehensive and nuanced understanding of Word Sense Induction and Lexical Semantic Change Detection tasks. Also, it will be critical to explore novel

[23]Upon acceptance, we will release the source code necessary to replicate our experiments.

quantile	WiC-path	binarize	ude	mod-hyp	fill-diag	dev-SPR	test-SPR	test-ARI
Method: Chinese Whispers								
6	ALL	True	True	w: lin	True	0.900	0.975	0.683
8	ALL	True	True	w: log	False	0.862	0.900	0.690
6	MCL→ru	False	True	w: top	False	0.896	0.400	0.600
6	ALL	True	True	w: top	False	0.955	0.564	0.488
5	MCL→ru	False	True	w: log	False	0.837	1.000	0.547
Method: Correlation Clustering								
4	MCL→ru	True	True	-	False	0.928	0.975	0.692
1	ALL	False	False	-	True	0.914	0.900	0.793
5	MCL→ru	True	True	-	False	0.901	0.700	0.572
5	MCL→ru	True	True	-	True	0.934	0.600	0.540
1	ALL	False	False	-	True	0.936	0.400	0.735
Method: WSBM								
6	ALL	True	True	geometric	True	0.926	0.975	0.673
5	ALL	True	True	poisson	False	0.892	0.900	0.726
6	ALL	True	True	geometric	False	0.939	0.700	0.517
0	enMCL	True	True	geometric	True	0.955	0.564	0.356
6	ALL	True	True	geometric	True	0.939	1.000	0.794
Method: Spectral Clustering/s								
4	ALL	False	True	-	True	0.940	0.975	0.753
4	ALL	False	True	-	True	0.939	0.999	0.818
3	MCL→ru	False	True	-	True	0.973	0.700	0.686
1	enMCL	False	True	-	True	0.962	0.700	0.619
2	MCL→es	False	True	-	True	0.946	0.800	0.701
Method: Spectral Clustering/c								
3	enMCL	True	True	-	False	0.928	0.872	0.758
0	ALL	True	True	-	True	0.944	0.900	0.807
0	ALL	True	True	-	True	0.936	0.700	0.651
4	enMCL	False	True	-	True	0.961	0.600	0.737
2	MCL→ru	False	True	-	True	0.958	0.800	0.560
Method: Spectral Clustering/e								
4	MCL→ru	True	True	-	True	0.833	0.516	0.317
4	MCL	False	True	-	True	0.840	0.999	0.833
7	enMCL	True	True	-	False	0.899	0.600	0.405
0	enMCL	True	True	-	True	0.944	0.462	0.299
7	MCL→es	False	True	-	False	0.897	0.400	0.278

Table 10: Detailed hyperparameters employed in constructing the WUG and the models during cross-validation, outlined for each fold and model within the context of the LSCD task. Each row corresponding to a method signifies a distinct fold.

techniques to enhance the adaptability of models for the LSCD task, ensuring they can efficiently handle both WSI and LSCD tasks with optimal performance taking into account the temporal aspects of target words.

Declarations

Funding

Frank D. Zamora-Reina and Felipe Bravo-Marquez were supported by ANID Millennium Science Initiative Program Code ICN17_002 and the National Center for Artificial Intelligence CENIA FB210017, Basal ANID. Dominik Schlechtweg has been funded by the project 'Towards Computational Lexical Semantic Change Detection' supported by the Swedish Research Council (2019–2022; contract 2018-01184) and by the research program 'Change is Key!' supported by Riksbankens Jubileumsfond (under reference number M21-0021). Nikolay Arefyev has received funding from the European Union's Horizon Europe research and innovation program under Grant agreement No 101070350 (HPLT).

Competing interests

The authors have no competing interests to declare.

Appendix A Experiments with models based on substitution

We also considered adding other architectures into our benchmark such as Amrami and Goldberg [76]. The author's work uses a model to generate substitutes given a list of target words, these substitutes are assigned to a set of dynamic clusters that generate predictions for a test dataset. We use this model to generate a baseline on our data, but we do not report its results due to its poor performance on our data. These results may be due to several factors: this method was originally used for data in English, while our data is in German. Similarly, another factor contributing to the suboptimal results is the sensitivity of standard clustering algorithms with off-the-shelf BERT to historical spelling variations, which can also be compounded by grammatical variations in some words [17].[24]

Appendix B Annotation Guidelines

Einführung.

Ihre Aufgabe ist es, Wortverwendungen einer Bedeutungsbeschreibung zuzuordnen. Ihnen werden Sätze wie in (1) vorgelegt, für die Sie die Bedeutung des markierten Zielworts, hier *Affentheater*, einer Bedeutungsbeschreibung wie in (2a) und (2b) zuweisen sollen.

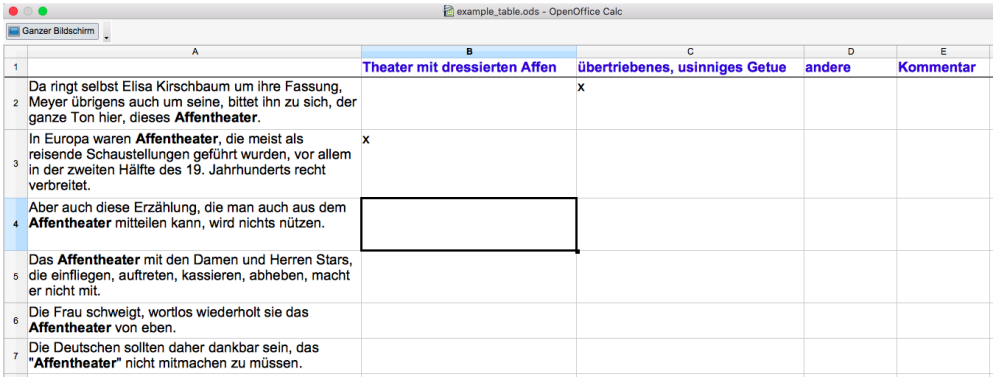
(B1) Da ringt selbst Elisa Kirschbaum um ihre Fassung, Meyer übrigens auch um seine, bittet ihn zu sich, der ganze Ton hier, dieses **Affentheater**.

(B2) a. Theater mit dressierten Affen
b. übertriebenes, usinniges Getue

Aufgabenstruktur.

Sie bekommen ein ODS-Tabellendokument wie in Tabelle B1 illustriert. Eine Zeile der Tabelle entspricht einem Satz. Die Spalten entsprechen Bedeutungsbeschreibungen für das jeweilige Zielwort. Das Zielwort ist in jedem Satz fett markiert. Ihre Aufgabe ist es, für jeden Satz die Bedeutungsbeschreibung zu markieren, welche am besten die Bedeutung des Zielworts im jeweiligen Satz beschreibt. Falls keine der Beschreibungen zutrifft, haben Sie die Möglichkeit, die Spalte "andere" zu markieren. Wenn Sie keine Entscheidung treffen können, markieren Sie bitte keine der Bedeutungsbeschreibungen und tragen nur eine Begründung in die Spalte "Kommentar" ein. Bitte wählen Sie nicht mehr als eine Bedeutungsbeschreibung aus und markieren Sie diese mit einem "x" in der jeweiligen Spalte.

[24] Ansell et al. [77] is another state-of-the-art model in the WSI task, we didn't it test it on our data because the model is trained on texts in English and it is too expensive to train it for other languages.



	A	B	C	D	E
1		Theater mit dressierten Affen	übertriebenes, usinniges Getue	andere	Kommentar
2	Da ringt selbst Elisa Kirschbaum um ihre Fassung, Meyer übrigens auch um seine, bittet ihn zu sich, der ganze Ton hier, dieses **Affentheater**.		x		
3	In Europa waren **Affentheater**, die meist als reisende Schaustellungen geführt wurden, vor allem in der zweiten Hälfte des 19. Jahrhunderts recht verbreitet.	x			
4	Aber auch diese Erzählung, die man auch aus dem **Affentheater** mitteilen kann, wird nichts nützen.				
5	Das **Affentheater** mit den Damen und Herren Stars, die einfliegen, auftreten, kassieren, abheben, macht er nicht mit.				
6	Die Frau schweigt, wortlos wiederholt sie das **Affentheater** von eben.				
7	Die Deutschen sollten daher dankbar sein, das **"Affentheater"** nicht mitmachen zu müssen.				

Table B1: Annotationstabelle.

Historische Sprachdaten.

Die Sätze für diese Annotationsaufgabe wurden aus historischen Korpora ausgelesen. Da sich Sprache mit der Zeit verändert, kann es sein, dass Worte anders benutzt werden, als Sie es gewohnt sind. Wenn Sie sich unsicher über die Bedeutung eines Wortes oder einer Konstruktion in einem Satz sind, versuchen Sie sie aus der Bedeutung des Kontexts zu erschließen. Die Sätze können sehr kurz oder sehr lang sein und ungrammatisch erscheinen. Außerdem können Worte aufgrund älterer Orthographie anders geschrieben sein, als Sie es gewohnt sind. Zudem wurden einige Buchstaben bei der automatischen Texterkennung eingescannter Dokumente falsch erkannt. Es wurde versucht, die Leserlichkeit durch Normalisierung spezieller Buchstaben zu moderner Orthographie zu verbessern. Versuchen Sie, diese Umstände zu ignorieren; konzentrieren Sie sich nur auf die Bedeutung des Zielwortes in seinem Kontext. Wenn Sie einen Satz zu fehlerhaft finden, um ihn zu verstehen, die Verwendung des Zielwortes mehrdeutig ist, oder die beiden Verwendungen des Zielwortes nicht zusammenpassen (d. h., nicht dasselbe Lemma haben), notieren Sie dies bitte auch in der Kommentarspalte.

Durchführung.

Während der Annotation der Sätze können Sie immer zu vorherigen Bewertungen zurückgehen und diese ändern; z. B. falls Sie Ihre Meinung ändern, nachdem Sie mehr Informationen bekommen haben.

Sie müssen nicht die ganze Datei in einer Sitzung annotieren. Wenn Sie einen Kommentar hinterlassen wollen, können Sie diesen in das Kommentarfeld eintragen.

Es kann hilfreich sein, die Rechtschreibprüfung zu deaktivieren, um nicht durch zusätzliche Hervorhebungen gestört zu werden.

Abschluss.

Bitte stellen Sie sicher, dass Sie nichts in der Datei ändern außer Spaltenbreite, Schriftgröße, Ihren Bewertungen und Kommentaren. Schicken Sie das annotierte Dokument an `schlecdk@ims.uni-stuttgart.de`. Wenn Sie noch Fragen zur Aufgabe haben, zögern Sie nicht, diese zu stellen.

References

[1] Langone, H., Haskell, B.R., Miller, G.A.: Annotating wordnet. In: Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL, Boston, MA, USA (2004)

[2] Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., Weischedel, R.: Ontonotes: The 90% solution. In: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. NAACL-Short '06, pp. 57–60. Association for Computational Linguistics, USA (2006)

[3] Erk, K., McCarthy, D., Gaylord, N.: Measuring word meaning in context. Computational Linguistics **39**(3), 511–554 (2013)

[4] Weaver, W.: Translation. In: Locke, W.N., Boothe, A.D. (eds.) Machine Translation of Languages, pp. 15–23. MIT Press, Cambridge, MA (1949/1955). Reprinted from a memorandum written by Weaver in 1949.

[5] Schütze, H.: Automatic word sense discrimination. Computational Linguistics **24**(1), 97–123 (1998)

[6] Blank, A.: Prinzipien des Lexikalischen Bedeutungswandels Am Beispiel der Romanischen Sprachen, p. 533. Niemeyer, Tübingen (1997)

[7] Tahmasebi, N., Dubossarsky, H.: Computational modeling of semantic change (2023)

[8] Paul, H.: Deutsches Wörterbuch: Bedeutungsgeschichte und Aufbau Unseres Wortschatzes, 10. edn. Niemeyer, Tübingen (2002)

[9] OED: Oxford English Dictionary. Oxford University Press, ??? (2009)

[10] Hamilton, W.L., Leskovec, J., Jurafsky, D.: Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, pp. 2116–2121 (2016)

[11] Ceron, T., Blokker, N., Padó, S.: Optimizing text representations to capture (dis)similarity between political parties. In: Proceedings of the 26th Conference on Computational Natural Language Learning. Association for Computational Linguistics, ??? (2022)

[12] Tahmasebi, N., Risse, T.: Finding individual word sense changes and their delay in appearance. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, Varna, Bulgaria, pp. 741–749 (2017)

[13] Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., Tahmasebi, N.: SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In:

Proceedings of the 14th International Workshop on Semantic Evaluation. Association for Computational Linguistics, Barcelona, Spain (2020). https://www.aclweb.org/anthology/2020.semeval-1.1/

[14] Rachinskiy, M., Arefyev, N.: GlossReader at LSCDiscovery: Train to select a proper gloss in english – discover lexical semantic change in spanish. In: Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change. Association for Computational Linguistics, Dublin, Ireland (2022)

[15] Teodorescu, D., Ohe, S., Kondrak, G.: UAlberta at LSCDiscovery: Lexical semantic change detection via word sense disambiguation. In: Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change. Association for Computational Linguistics, Dublin, Ireland (2022)

[16] Giulianelli, M., del Tredici, M., Fernández, R.: Analysing lexical semantic change with contextualised word representations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3960–3973. Association for Computational Linguistics, Online (2020)

[17] Laicher, S., Kurtyigit, S., Schlechtweg, D., Kuhn, J., Walde, S.: Explaining and Improving BERT Performance on Lexical Semantic Change Detection. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 192–202. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.eacl-srw.25 . https://aclanthology.org/2021.eacl-srw.25

[18] Homskiy, D., Arefyev, N.: DeepMistake at LSCDiscovery: Can a multilingual word-in-context model replace human annotators? In: Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change. Association for Computational Linguistics, Dublin, Ireland (2022)

[19] Zamora-Reina, F.D., Bravo-Marquez, F., Schlechtweg, D.: LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In: Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change. Association for Computational Linguistics, Dublin, Ireland (2022). https://aclanthology.org/2022.lchange-1.16/

[20] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423

[21] Schlechtweg, D.: Human and computational measurement of lexical semantic change. PhD thesis, University of Stuttgart, Stuttgart, Germany (2023). http:

//dx.doi.org/10.18419/opus-12833

[22] Davletov, A., Arefyev, N., Gordeev, D., Rey, A.: LIORI at SemEval-2021 task 2: Span prediction and binary classification approaches to word-in-context disambiguation. In: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pp. 780–786. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.semeval-1.103 . https://aclanthology.org/2021.semeval-1.103

[23] Navigli, R.: Word sense disambiguation: a survey. ACM Computing Surveys **41**(2), 1–69 (2009)

[24] Tuggy, D.: Ambiguity, polysemy, and vagueness. Cognitive Linguistics **4**(3), 273–290 (1993)

[25] Cruse, D.A.: 2. In: Saint-Dizier, P., Viegas, E. (eds.) Polysemy and related phenomena from a cognitive linguistic viewpoint. Studies in Natural Language Processing, pp. 33–49. Cambridge University Press, ??? (1995). https://doi.org/10.1017/CBO9780511527227.004

[26] Kilgarriff, A.: "I don't believe in word senses". Computers and the Humanities **31(2)** (1997)

[27] Hanks, P.: Do word meanings exist? Computers and the Humanities **34**(1/2), 205–215 (2000)

[28] Palmer, M., Dang, H., Fellbaum, C.: Making fine-grained and coarse-grained sense distinctions, both manually and automatically. Natural Language Engineering **13**, 137–163 (2007) https://doi.org/10.1017/S135132490500402X

[29] Schlechtweg, D., Schulte im Walde, S., Eckmann, S.: 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, New Orleans, Louisiana, pp. 169–174 (2018). https://www.aclweb.org/anthology/N18-2027/

[30] McCarthy, D., Navigli, R.: SemEval-2007 task 10: English lexical substitution task. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), pp. 48–53. Association for Computational Linguistics, Prague, Czech Republic (2007). https://aclanthology.org/S07-1009

[31] McCarthy, D., Apidianaki, M., Erk, K.: Word sense clustering and clusterability. Computational Linguistics **42**(2), 245–275 (2016)

[32] Schlechtweg, D., Tahmasebi, N., Hengchen, S., Dubossarsky, H., McGillivray,

B.: DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 7079–7091. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https://aclanthology.org/2021.emnlp-main.567

[33] Schlechtweg, D., Castaneda, E., Kuhn, J., Schulte im Walde, S.: Modeling sense structure in word usage graphs with the weighted stochastic block model. In: Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics, pp. 241–251. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.starsem-1.23 . https://aclanthology.org/2021.starsem-1.23

[34] Kutuzov, A., Øvrelid, L., Szymanski, T., Velldal, E.: Diachronic word embeddings and semantic shifts: a survey. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1384–1397. Association for Computational Linguistics, Santa Fe, New Mexico, USA (2018)

[35] Cook, P., Lau, J.H., McCarthy, D., Baldwin, T.: Novel word-sense identification. In: 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, Dublin, Ireland, pp. 1624–1635 (2014)

[36] Rodina, J., Kutuzov, A.: RuSemShift: a dataset of historical lexical semantic change in Russian. In: Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020). Association for Computational Linguistics, ??? (2020)

[37] Kutuzov, A., Touileb, S., Mæhlum, P., Enstad, T., Wittemann, A.: NorDiaChange: Diachronic semantic change dataset for Norwegian. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 2563–2572. European Language Resources Association, Marseille, France (2022). https://aclanthology.org/2022.lrec-1.274

[38] Chen, J., Chersoni, E., Schlechtweg, D., Prokic, J., Huang, C.-R.: ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection. In: Proceedings of the 4th International Workshop on Computational Approaches to Historical Language Change. Association for Computational Linguistics, Singapore (2023). https://aclanthology.org/2023.lchange-1.10/

[39] Perrone, V., Palma, M., Hengchen, S., Vatri, A., Smith, J.Q., McGillivray, B.: GASC: Genre-aware semantic change for ancient Greek. In: Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pp. 56–66. Association for Computational Linguistics, Florence, Italy (2019)

[40] Basile, P., Caputo, A., Caselli, T., Cassotti, P., Varvara, R.: Overview of the EVALITA 2020 Diachronic Lexical Semantics (DIACR-Ita) Task. In: Basile, V.,

Croce, D., Di Maro, M., Passaro, L.C. (eds.) Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020). CEUR.org, Online (2020)

[41] McGillivray, B.: Latin lexical semantic annotation. Literary and Linguistic Data Service (2021). http://hdl.handle.net/20.500.14106/2568

[42] Asahara, M., Ikegami, N., Suzuki, T., Ichimura, T., Kondo, A., Kato, S., Yamazaki, M.: CHJ-WLSP: Annotation of 'word list by semantic principles' labels for the corpus of historical Japanese. In: Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages, pp. 31–37. European Language Resources Association, Marseille, France (2022). https://aclanthology.org/2022.lt4hala-1.5

[43] DWDS: Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart, hrsg. v. d. Berlin-Brandenburgischen Akademie der Wissenschaften. https://www.dwds.de/. Accessed: 02.02.2021 (2021)

[44] Manjavacas Arevalo, E., Fonteyn, L.: Non-parametric word sense disambiguation for historical languages. In: Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities, pp. 123–134. Association for Computational Linguistics, Taipei, Taiwan (2022). https://aclanthology.org/2022.nlp4dh-1.16

[45] Pilehvar, M.T., Camacho-Collados, J.: WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 1267–1273. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1128

[46] Camacho-Collados, J., Pilehvar, M.T.: From word to sense embeddings: A survey on vector representations of meaning. Journal of Artificial Intelligence Research **63**, 743–788 (2018)

[47] Armendariz, C.S., Purver, M., Ulčar, M., Pollak, S., Ljubešić, N., Granroth-Wilding, M.: CoSimLex: A resource for evaluating graded word similarity in context. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 5878–5886. European Language Resources Association, Marseille, France (2020). https://aclanthology.org/2020.lrec-1.720

[48] Kutuzov, A., Pivovarova, L.: Rushifteval: a shared task on semantic shift detection for russian. Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference (2021)

[49] Arefyev, N., Rachinskiy, M.: Zero-shot cross-lingual transfer of a gloss language

model for semantic change detection, vol. 2021-June, pp. 578–586 (2021). https://doi.org/10.28995/2075-7182-2021-20-578-586

[50] Schlechtweg, D., Virk, S.M., Arefyev, N.: The LSCD Benchmark: a Testbed for Diachronic Word Meaning Tasks (2024)

[51] Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Comput. Linguist. **34**(4), 555–596 (2008) https://doi.org/10.1162/coli.07-034-R2

[52] Jurgens, D., Klapaftis, I.: Semeval-2013 task 13: Word sense induction for graded and non-graded senses. In: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 290–299 (2013)

[53] Navigli, R.: A quick tour of word sense disambiguation, induction and related approaches. In: SOFSEM 2012: Theory and Practice of Computer Science: 38th Conference on Current Trends in Theory and Practice of Computer Science, ŠpindlerŮv Mlỳn, Czech Republic, January 21-27, 2012. Proceedings 38, pp. 115–129 (2012). Springer

[54] Raganato, A., Camacho-Collados, J., Navigli, R., *et al.*: Word sense disambiguation: a uinified evaluation framework and empirical comparison. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, vol. 1, pp. 99–110 (2017)

[55] Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification **2**, 193–218 (1985)

[56] Spearman, C.: The proof and measurement of association between two things. American Journal of Psychology **15**, 88–103 (1904)

[57] Arefyev, N., Fedoseev, M., Protasov, V., Homskiy, D., Davletov, A., Panchenko, A.: Deepmistake: Which senses are hard to distinguish for a word-in-context model, vol. 2021-June, pp. 16–30 (2021)

[58] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.747 . https://aclanthology.org/2020.acl-main.747

[59] Martelli, F., Kalach, N., Tola, G., Navigli, R.: Semeval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (mcl-wic). In: SEMEVAL (2021)

[60] Pasini, T., Raganato, A., Navigli, R.: XL-WSD: An extra-large and cross-lingual

evaluation framework for word sense disambiguation. In: Proc. of AAAI (2021)

[61] Biemann, C.: Chinese whispers: An efficient graph clustering algorithm and its application to natural language processing problems. In: Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing. TextGraphs-1, pp. 73–80. Association for Computational Linguistics, USA (2006)

[62] Ustalov, D., Panchenko, A., Biemann, C., Ponzetto, S.P.: Watset: Local-Global Graph Clustering with Applications in Sense and Frame Induction. Computational Linguistics **45**(3), 423–479 (2019) https://doi.org/10.1162/COLI_a_00354

[63] Bansal, N., Blum, A., Chawla, S.: Correlation clustering, vol. 56, pp. 238–247 (2002). https://doi.org/10.1109/SFCS.2002.1181947

[64] Kurtyigit, S., Park, M., Schlechtweg, D., Kuhn, J., Walde, S.: Lexical Semantic Change Discovery. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Online (2021). https://aclanthology.org/2021.acl-long.543/

[65] Baldissin, G., Schlechtweg, D., Schulte im Walde, S.: DiaWUG: A Dataset for Diatopic Lexical Semantic Variation in Spanish. In: Proceedings of the 13th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France (2022). https://aclanthology.org/2022.lrec-1.278/

[66] Aksenova, A., Gavrishina, E., Rykov, E., Kutuzov, A.: RuDSI: graph-based word sense induction dataset for Russian. arXiv (2022). https://doi.org/10.48550/ARXIV.2209.13750 . https://arxiv.org/abs/2209.13750

[67] Aicher, C., Jacobs, A.Z., Clauset, A.: Learning latent block structure in weighted networks. Journal of Complex Networks **3**(2), 221–248 (2014) https://doi.org/10.1093/comnet/cnu026

[68] Peixoto, T.P.: Nonparametric weighted stochastic block models. Physical Review E **97** (2017) https://doi.org/10.1103/PhysRevE.97.012306

[69] Kotchourko, S.: Optimizing Human Annotation of Word Usage Graphs in a Realistic Simulation Environment. Bachelor thesis

[70] Tunc, B.: Optimierung Von Clustering Von Wortverwendungsgraphen. https://elib.uni-stuttgart.de/handle/11682/11923

[71] Luxburg, U.: A tutorial on spectral clustering (2007) https://doi.org/10.48550/ARXIV.0711.0189

[72] Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20**, 53–65

(1987) https://doi.org/10.1016/0377-0427(87)90125-7

[73] Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics **3**(1), 1–27 (1974) https://doi.org/10.1080/03610927408827101 https://www.tandfonline.com/doi/pdf/10.1080/03610927408827101

[74] Cawley, G.C., Talbot, N.L.: On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research **11**, 2079–2107 (2010)

[75] Kutuzov, A., Pivovarova, L.: Three-part diachronic semantic change dataset for Russian (2021)

[76] Amrami, A., Goldberg, Y.: Towards better substitution-based word sense induction. arXiv preprint arXiv:1905.12598 (2019)

[77] Ansell, A., Bravo-Marquez, F., Pfahringer, B.: Polylm: Learning about polysemy through language modeling. arXiv preprint arXiv:2101.10448 (2021)