Annotating Word Meanings Over Time: The Trade-off Between Scalability, Reliability and Expressivity Power

Pierluigi Cassotti, Nina Tahmasebi
University of Gothenburg
{pierluigi.cassotti,nina.tahmasebi}@gu.se

Abstract

Annotating the meanings of a word over time in order to document their emergence or disappearance presents substantial implementation challenges. These difficulties arise for several reasons, notably the need for sufficient expressive power in the annotation paradigm to capture unconventional or rare meanings, as well as issues of scalability related to the number of annotations required. The first challenge is particularly acute in the context of historical texts, where modern annotators must interpret word meanings in sources that are temporally distant and often absent from contemporary dictionaries and language use. The second challenge is inherent to the distribution of word meanings, which tend to occur sparsely and intermittently over long time spans. In this paper, we examine several annotation paradigms, discussing their respective advantages and limitations. We also present a pilot study on English and Swedish. Our results indicate that a usage-sense inventory based annotation paradigm can be adopted in place of a usage-pairs-based approach while maintaining expressivity power and reducing the complexity from quadratic to linear.

Keywords: word sense disambiguation, lexical semantics, diachronic linguistics

1. Introduction

The creation of datasets capturing word meanings by means of human annotation has long been, and continues to be, fundamental to the development of language technologies capable of understanding human language and its dynamics. However, a persistent challenge in annotation is that word meanings are latent, fuzzy variables whose boundaries are difficult to define. There is rarely a clear point at which one meaning ends and another begins, neither within a single period nor across time periods. Often, the set of meanings present in text vary across sources; different genre will use words in different ways (e.g., think of a *worm* in magazines on gardens or computers).

As a result of the difficulty to capture meanings, researchers commonly rely on an approximation that equates meanings with *senses* by which we refer to word meanings that have been defined by expert linguists or lexicographers as discrete, tangible entities, such as the sense definitions found in dictionaries. In the annotation task, people then get to label words in naturally occurring text with one (or more) senses from the sense repository. A task often called word sense disambiguation. As with most standard annotation tasks, each instance is annotated by several annotators and the majority label is chosen.

But even using a sense repository as an approximation, significant challenges remain. One such challenge is that the sense inventories, although created by experts, reflect the experts' conceptualizations of meaning. This means that the way in which the semantic space is divided is not necessarily deterministic, nor are the resulting senses fully reflected in natural language. Indeed, even among experts, such as lexicographers, there is often disagreement about how meanings should be delineated. This issue is commonly referred to as the lumpers versus splitters problem where on the one hand, lumpers tend to group similar meanings together, and on the other, splitters aim to document even the smallest variations. When a dictionary is used as a reference for sense inventories, this dichotomy is reflected in the representation of word meanings. And because dictionaries are typically the product of the work of multiple teams, compiled over long periods of time and subject to updates and revisions, the final resource is often inherently heterogeneous affecting the resulting annotations.

Also on the computational models, this has an effect. Models trained and evaluated on such resources tend to be effective at recognizing clear semantic distinctions, such as homonymy or cases of polysemy involving evident shifts in domain, for example in metaphorical extensions of meaning. However, when the same models are required to recognize subtle differences in meaning, their performance deteriorates significantly.[1]

An alternative way of annotating for word meaning (both synchronically and diachronically), has gained popularity in the past decade. Instead of basing itself in the use of sense repositories from external sources, it makes no assumptions on which meanings exist for a word in the corpus. The modeling paradigm instead compares pairs of usages

[1] It should be noted that even humans tend to show lower agreement in such cases, suggesting that the underlying reason is likely that individual perception plays a large role. Consequently, the internal models of meaning that each individual holds become more influential in these instances.

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 88–98
12 May 2026.

of a word sampled from the corpus and allows the user to make decisions on the similarity of a target word across two usages. Based on judgments on a larger set of usage pairs, senses can be induced. Such datasets exist for many languages under the name Word-In-Context (WiC) which refers to synchronic data with binary annotation (same meaning/ different meaning), or (Diachronic) Word Usage Graphs (DWUG). In the WiC case, the aim is often to use the data for evaluation of large language models and hence to have many words with few usage pairs for each. In the DWUG case, however, the aim is to study words and their meanings over time, and thus many more usage pairs are included and consequently, many more annotations are needed.

In fact, when we want to annotate word meanings *over time*, as opposed to traditional word sense disambiguation that takes place synchronically, we must take into account the volume of data needed to answer the question: when did a meaning appear, change, or disappear? This necessitates sampling from large scale historical archives, frequently involving thousands of occurrences of a word for each time period to reduce the number of annotations needed, while maintaining the distribution of the senses present in the text. Because of the scalability issue, often a gross simplification has been used: two far apart time periods are compared to each other to determine change in the senses found for a word. However, this view is very limited and does not allow us to build realistic models that can trace the dynamics of individual senses over time. But, when we take further time periods into account, the scale of annotation explodes and becomes infeasible.

In this paper, we address the challenge of annotation scalability through a pilot study in which we compare different annotation paradigms. Based on our results, we argue that a more efficient approach can be adopted while maintaining a level of quality comparable to that of existing methods. These results will then feed into our large-scale annotations of meaning change across multiple time periods.

2. Annotation Paradigms

In this section, we describe the different paradigms for annotating word meanings and introduce the terminology used throughout the paper. A usage of a word refers to a specific instance of that word in-context, that is, a word instantiated within a naturally occurring sentence. A sense is a discrete entity that captures a particular interpretation of a word, and a sense inventory is the set of all possible senses associated with a word. In this article, references to definitions generally denote those provided in dictionaries. When we discuss scalability, these have to be multiplied by the number of desired annotations per instance. Standard is to have three annotators and take as the majority vote as the label in the case of binary labels, and an average of the (ordinal) values otherwise.

Example of Annotation We use an example to illustrate, in numerical terms, how many annotations are required. In our example, we consider a specific word, with 3 senses, 20 usages per time period (the minimum adopted so far in (Zamora-Reina et al., 2022)), and 10 time points.

2.1. Usage-Usage (U-U)

This paradigm follows a two step procedure for annotating word meanings. Annotators are presented with a single pair of word uses at the time, and are asked to assess how similar these uses are, that is, how similar the meaning of the word is across the two contexts. As an example, we can use the target word *rock* in the sentences *I listened to rock.* and *I threw a rock.* where the similarity is low, while *I listened to rock.* and *I went to a rock concert.* has a high similarity. The resulting similarity judgments across multiple pairs of usages for the same target word are then aggregated and can be thought of as a graph. In each graph representing a single target word, nodes are the usages of the target word while edges between the nodes correspond to the similarity judgments. The edges of the graph group together usages that are most similar to one another, where each group can be said to correspond to a distinct meaning of the target word.

Similarity between pairs of uses can be assessed either in a binary manner or on a continuous scale. In the binary setting, corresponding to the Word in Context task, annotators choose between two options: either the two uses express the same meaning or they express different meanings. When similarity is expressed on a graded scale, the task is often referred to as a Graded Word in Context task.

An alternative to binary judgments is the semantic relatedness scale introduced in the DUREL framework, which consists of four values ranging from 1 to 4. Here, 1 denotes Unrelated, 2 Distantly Related, 3 Closely Related, and 4 Identical. This scale is inspired by Blank's notion of semantic proximity and establishes a correspondence between the four values and the categories of homonymy, polysemy, context variance, and identity.

Scalability: This annotation paradigm scales **quadratically** since each usage has to be compared to each other usage. In diachronic annotation, this becomes problematic. Consider our example, one word with 20 usages for 10 time points, we have a total of 200 usages. The total number of possible pairwise comparisons (annotations)

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

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

would then be $(U(U-1))/2 =$ **19,900**. If each comparison is annotated by at least three annotators, this amounts to 59,700 individual annotations. In practice however, it is infeasible to annotate each instance with every other (and has only been done for ChiWUG where 40 Chinese words with 40 usages each were annotated by two annotators, resulting in about 60,000 annotations (Chen et al., 2023a)). Instead, common practice is to sample among the usage pairs while ensuring that sufficient connectivity remains among the usages, see e.g., (Schlechtweg et al., 2021a).

Previous annotations and experiences Most notably, this style of annotation was used for the SemEval-2020 Task 1 on Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020). The work was followed up by benchmarks for multiple languages, including Russian (Kutuzov and Pivovarova, 2021a,b), Spanish (Zamora-Reina et al., 2022), Norwegian (Kutuzov et al., 2022), and Chinese (Chen et al., 2023b, 2022), as well as reannotations for Swedish, German and English with smaller set of usages and higher density of usage pair annotations (Schlechtweg et al., 2024). For binary Usage-Usage annotations with a temporal dimension, also TempoWiC (Loureiro et al., 2022) is available, focusing on shorter time spans within social media. In general, all WiC data where the labels are assigned by human annotators (as opposed to derived from external resources e.g. WordNet) fall in this category, see e.g., (Cassotti et al., 2023).

In terms of temporal bins, at least for the original SemEval corpora, each subcorpus was around 40-50 years in size[2] and thus significant amount of change occurred already in within a single time period. With limited annotation budgets, there is a natural trade-off between a high number of usages (regardless of time period) versus a high density of annotation between the usage pairs. If we can afford a limited set of annotations, the more usages we have, the less dense becomes the graph. For 200 usages, there are 19,900 pairs. A 5,000-annotation budget will result in a density of 0.25. While for a smaller set of 100 usages, 5,000 annotations covers the whole graph. This trade-off is discussed by Schlechtweg et al. (2024) who showed that it is preferable to have a higher degree of annotated pairs rather than a larger number of usages, as it leads to more robust sense clusters.

[2]This choice was made for several reasons. Firstly, corpora tend to grow in size over time. So to ensure that a sufficient number of usages exist for most words included in the benchmark, the initial time period needed to be broad. Secondly, the choice to have large time bins for each subcorpus was made to increase representation of each sense.

2.2. Usage-Sense (U-S)

In this paradigm, annotators are presented with a word usage and the definition of one of its senses, and they are asked to determine whether the sense definition corresponds to the meaning expressed in that particular usage. As before, this judgment can be made either in a binary fashion or on a continuous scale. In the binary setting, annotators decide whether the use and the definition reflect the same meaning or not. This task is commonly referred to as Target Sense Verification (TSV). When a continuous scale is employed, annotators assess how similar the meaning expressed by the word use is to the meaning described by the proposed sense definition, typically using the four value semantic relatedness scale.

We included this annotation paradigm in our pilot to allow for each sense to be considered independently from the other senses, as to avoid potential bias where users do not assign a sense because another one is a better fit. Another advantage is that users do not need to be overwhelmed with information, making the task complex. If a word has 20 senses as opposed to three, the choice of which one fits the best becomes much more difficult. If instead the annotator sees a single sense at the time, the choice is independent of the number of senses a word has.

Scalability: This annotation paradigm scales **linearly** with the number of usages. For each word, we will have kU instances to annotate, where k is the number of senses of the target word. In our example, this results in **600 annotations** (each of the 200 usages is paired with each of the 3 senses). That is two orders of magnitude less the than usage-usage annotation framework.

Previous annotations and experiences: To the best of our knowledge, this annotation schema has not been tested in previous research as the obvious downside is an increasing annotation need without obvious benefits. Erk et al. (2013) and Cassotti and Tahmasebi (2025), however follow this paradigm to some extent as they ask annotators to provide for a usage, a rating with respect to each sense in the sense inventory. While these are not independent, and can be randomized and mixed with annotations for other usages to maximize anno-

Lemma	Sense Definitions
graft	**1)** A shoot or scion inserted in a groove or slit made in another stock, [..] **2)** Surgery. 'A portion of living tissue transplanted [..] **3)** A ditch; a moat; [..] **4)** The depth of earth that may be thrown up at once with a spade; [..] **5)** A kind of spade, used in digging drains. **6)** Work, esp. hard work. / A trade, craft. **7)** The obtaining of profit or advantage by dishonest or shady means; [..]
twist	**1)** A divided object or part. **2)** The twisting of threads into a cord, and derived senses. **3)** A slight or weak support upon which something depends; [..] **4)** An act or the action of turning on or as on an axis; [..] **5)** A dance in which the body is twisted from side to side; [..] **6)** A young woman, a girl. [..]
konduktör	**1)** Person som säljer och kontrollerar biljetter och andra färdbevis på tåg. **2)** Elektroteknisk ledare, överförare av kraft, [..] Kan också användas bildligt. **3)** Titel för arbetsledare eller uppsyningsman vid slott, [..] **4)** En i militärutbildad byggnads- eller arbetsledare vid fästning och dylikt; [..] **5)** Arbetsledare vid byggnadsverk [..]
motiv	**1)** Underliggande orsak till viss handling **2)** Ämne för konstnärlig framställning särskilt inom bildkonst och litteratur; [..] **3)** Minsta melodisk-rytmiska enhet av musikstycke, vanligen återkommande och lätt igenkännlig **4)** Spets- eller broderiarbete i avpassade delar till påsättning på underkläder, [..]

Table 2: U-S and U-SI sense inventories for graft, twist, konduktör, and motiv.

tator objectivity, the results are as close as currently exist.

2.3. Usage-Sense Inventory (U-SI)

In this paradigm, annotators are presented with a word usage and the definitions of all the word's senses at once. It follows a standard word sense disambiguation setting, where the annotators choose either the best fit, or alternatively, the set of best fitting senses. Like for prior annotation paradigms, this judgment can be made either in a binary fashion or on a continuous scale. Prior research has shown that there are great benefits to allowing annotators to provide a graded rating to *every* sense, rather than choosing top senses. Erk et al. (2013) found that the latter can lead to a bias in assigning a single best sense (and by implication, not assigning the rest). However, grading all senses increases the number of annotations and becomes equivalent to the U-S scenario above.

Scalability: This annotation paradigm is the most scalable one (assuming one does not require a rating for each and every sense). The schema scales with the number of usages and is thus linear, as each usage is judged only once with respect to *all* senses (in our annotation example this corresponds to **200 annotations**). Despite the linear scaling, the annotation task becomes complex if each usage is to be judged with respect to a large set of senses. The cognitive effort for keeping all senses in memory to make a good choice increases with the number of senses. After annotating a larger set of usages, the annotator typically becomes familiar with the senses, which reduces the required effort. However, a cold-start problem remains if the annotator pauses for some time and later returns to the task. Thus, to have maximum gain from this annotation schema, it is important not to overload the user with too many, or too long, sense descriptions.

Previous annotations and experiences: This annotation paradigm is standard for word sense disambiguation tasks. While typically, in WSD, the temporal dimension is lacking, the task itself is identical. Prominent examples of WSD data are SemCor (Miller et al., 1994), Sense-Eval (Snyder and Palmer, 2004; Edmonds and Cotton, 2001a,b) and SemEval (Moro and Navigli, 2015; Navigli et al., 2013; Agirre and Soroa, 2007a). A notable difference between some of these corpora and the paradigm we are proposing above is that some of these corpora disambiguate not only a target word but all words in the sentence. We propose disambiguating only a target word.

2.4. Other Paradigms

Beyond the paradigms described above, prior work has explored bottom-up sense induction through context clustering (Agirre and Soroa, 2007b; Schütze, 1998), substitute-based approaches (McCarthy and Navigli, 2007), and direct semantic change judgments across time (Cook et al., 2014).

2.5. Reliability

In the U-U paradigm, previous work (Schlechtweg et al., 2021b; Cassotti and Tahmasebi, 2025) clearly shows that annotators struggle to reach agreement on adjacent categories (e.g., Identical vs. Closely Related). This error can be systematic, stemming from differing perceptions of semantic proximity among annotators, or it may arise from the intrinsic difficulty of the task. In particular, the scale values

are not anchored to easily interpretable or objective reference points, which can even lead to inconsistencies within the same annotator's judgments.

3. Expressivity Power

In general, U–U is a paradigm with greater expressive power than U–S and U–SI, because the annotation scheme does not impose explicit constraints on which meanings can be labeled, leaving the representation of meaning implicit. However, in practice this theoretical advantage does not fully materialize, due to the clustering choices.

Clustering is typically performed using correlation clustering (Bansal et al., 2004) with a fixed threshold to binarize the relations between nodes, commonly set at 2.5. Under this scheme, all node pairs with a score above the threshold are treated as valid links, while those below it are discarded. As a consequence, clusters may include node pairs with an average score of 3, grouping them together despite only moderate relatedness.

Regarding the senses used in U-S and U-SI paradigms, we can make several choices. The first one is to use all senses available at all time periods. This results in redundant senses presented to the user at some time periods, as, e.g., the sense of e.g., computer virus did not exist for the word *worm* in year 1900. However, to reduce the number of senses to only those presumed active at the time from which the sentence stems, has several consequences. Firstly, we would need to make assumptions about validity periods of senses on the basis of dictionaries. However, often dictionaries are normative and have poor correspondence with empirical data; words can be used with specific senses before the senses conventionalize and thus are attested in the dictionary. Secondly, the order of the sentences annotated should be random, also across time periods, to avoid that the annotators make a priori choices based on knowledge of the time period. The set of senses presented to the annotators can thus be a biasing factor and should therefore include all senses. Finally, from a cognitive perspective, we assume that the annotators get aquatinted with the senses and their order, thus changing the set, or order, of the senses between instances can lead to additional cognitive load.

4. Pilot

Our long-term goal is to annotate a set of words across 10 time periods. However, our limited budget constrains the total *annotation hours* we can fund, which in turn limits the number of annotations we can obtain. The higher the price of an annotation paradigm, the fewer words and/or usages we can afford to annotate. Before we scale up the annotations, we therefore need a good evaluation of the complexity of the different annotation paradigms together with an estimate of how many annotations per hour that can be done.

	t_1	t_2
English	CCOHA 1810–1860	CCOHA 1960–2010
Swedish	Kubhist 1790–1830	Kubhist 1895–1903

Table 3: Time periods of subcorpora for English and Swedish from which annotation data was sampled.

4.1. Data Annotation

To test the different annotation paradigms, we started from the existing resampled DWUG dataset for English and Swedish (Schlechtweg et al., 2024). In particular, we focused on two English target words (*graft* and *twist*) and two Swedish target words (*motiv* 'motif' and *konduktör* 'conductor'). The choice of languages, and thus dataset, was pragmatic as we plan on annotating both in our future work.

In the *resampled* DWUG dataset (DWUG henceforth), each of these four words includes up to 50 usages, one half drawn from period T1 and T2 respectively. The corresponding time spans and corpora are reported in Table 3. For each word there is a maximum of 1,225 usage pairs that can be annotated. However, only a subset of these pairs were actually annotated: on average, approximately 35% for English and 60% for Swedish. In total, the number of annotated usage pairs amounts to 855 for *motiv*, 981 for *konduktör*, 342 for *graft*, and 509 for *twist*.

We extended the DWUG dataset (U-U) by annotating the usages of the four target words, employing the U-S and U-SI paradigms[3]. For Swedish, we recruited six native Swedish annotators. For English, we selected three annotators: two native speakers of British English and one native speaker of Canadian English.[4] We recruited both sets of annotators such that they had studied either linguistics or languages at the university level. Swedish annotators were paid standard hourly wages for research assistants, while English annotators received a lump sum voucher.

Sense Inventories One of the challenges in both the U-S and U-SI paradigms is defining the set of senses to be used for annotation. In particular, for diachronic annotation it is essential to ensure that the sense inventory also includes senses that are now obsolete or have disappeared, as well as those that have been introduced more recently.

[3]The dataset and the code are available on Github

[4]Our English annotators were a convenience sample.

Dictionaries rarely capture all the senses of a word; in practice, it is often necessary to aggregate information from multiple lexicographic sources. Moreover, sense definitions should be as precise as possible, while minimizing overlap between definitions.

It is also necessary to adjust the granularity of the senses to the specific research objective. Some dictionaries adopt an extreme splitter approach, listing dozens of senses for a single word. This creates difficulties in both paradigms. In the U-S paradigm, the number of examples to annotate with additional senses. In the U-SI paradigm instead, the annotator's cognitive load increases, as they must select the correct sense from a large set of closely related definitions.

Given the sensitivity and complexity of constructing a sense inventory, we believe that the only way to ensure its reliability is through careful manual curation[5]. For this paper, a member of our team curated the sense inventories for the four target words, drawing on a combination of definitions from SO and SAOB for Swedish [6] and from the OED for English [7]. In some cases, the definitions were slightly modified to make them more general and to cover a broader semantic field. The definitions for the four words are reported in Table 2.

Annotation Setting For the annotation process, we used Qualtrics[8]. For each language, we created a survey including the two target words. The survey presents all examples sequentially, first for the U-S paradigm and then for the U-SI paradigm. For all annotators of a language, we use the same order of paradigms, words, and usages so that, when annotators begin with the U-S paradigm, they are not yet familiar with the complete sense inventory but instead discover them progressively. The annotation order was as follows:

- Swedish:
  Usage-Sense (konduktör) → Usage-Sense (motiv) → Usage-Sense Inventory (konduktör) → Usage-Sense Inventory (motiv)

- English:
  Usage-Sense (twist) → Usage-Sense (graft) → Usage-Sense Inventory (twist) → Usage-Sense Inventory (graft)

[5]Manual curation of the sense inventory also allows a hypothesis-driven approach: If the aim is to study certain aspects of change, then more fine-grained senses can be included to reflect those aspects while other senses can remain more coarse.

[6]https://svenska.se

[7]https://www.oed.com

[8]https://www.qualtrics.com

Word	Lang.	α(U-S)	α(U-SI)
konduktör	swedish	0.818	0.759
motiv	swedish	0.888	0.882
twist	english	0.552	0.593
graft	english	0.739	0.795

Table 4: **Inter-annotator agreement**. Krippendorff's α for Usage-Sense (U-S) and Usage-Sense Inventory (U-SI) annotation.

Given the high number of examples to be annotated, annotators were given two weeks to complete the entire survey. During this period, they were allowed to pause the annotation process and resume it later from where they had left off. We recorded several metadata variables, including the time elapsed between opening an example and the first click, the last click, and the final submission, as well as the total number of clicks.

In the U-S paradigm, annotators are provided with a word's usage and a single candidate sense of that word. They are asked to answer yes or no as to whether the definition corresponds to the meaning of the word in that specific usage. Annotators may also select *Cannot decide*. In that case, they must choose between two options: *I cannot decide because the text is noisy (e.g., OCR errors)* or *I cannot decide because the text is ambiguous*. This choice is exclusive: if annotators select Cannot decide, they cannot simultaneously answer yes or no.

In the U-SI paradigm, annotators are provided with a word in context and the full inventory of senses. They are also given the option to indicate that none of the provided senses adequately captures the meaning of the word (*Other*), in which case they must supply a textual explanation of the intended meaning.

Annotators are allowed to select *Other* simultaneously with one or more of the predefined senses. By contrast, the choice of Cannot decide (again divided into *noise* and *ambiguity*) remains exclusive in this paradigm as well. The annotation guidelines and the interface were provided in the annotators' native language.

4.2. U-S vs U-SI paradigm

Agreement To assess whether the U-S and U-SI paradigms yield comparable annotations, we examine the bootstrap estimates of Cohen's κ of the intra-annotator agreement reported in Tables 5 and 6. These estimates directly quantify cross-paradigm agreement at the annotator level, clustered by usage with 95% confidence intervals. The inter-annotator agreement instead is reported in Table 4

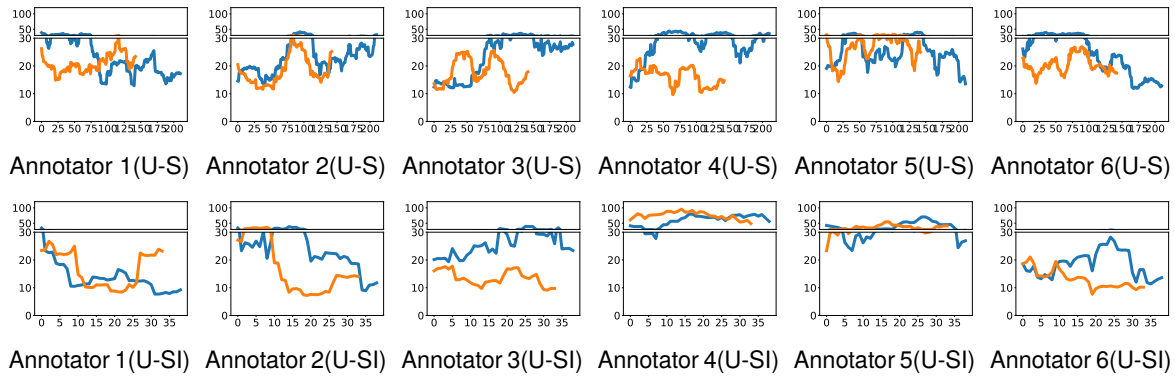

Figure 1: Time in seconds (y-axis) spent by each annotator (shown in the columns) and annotation paradigm (U-S first row, U-SI second row), relative to the number of annotated examples (x-axis) for the words konduktor and motiv.

	konduktör	motiv
ann1	1.000 [1.000, 1.000]	0.981 [0.935, 1.000]
ann2	0.853 [0.763, 0.929]	1.000 [1.000, 1.000]
ann3	0.797 [0.691, 0.892]	0.953 [0.892, 1.000]
ann4	0.891 [0.805, 0.956]	0.907 [0.812, 0.979]
ann5	0.762 [0.656, 0.860]	0.938 [0.880, 0.985]
ann6	0.946 [0.882, 1.000]	0.915 [0.809, 0.983]

Table 5: **Intra-annotator agreement.** Bootstrap estimates of Cohen's κ (U-S vs U-SI), clustered by sentence, with 95% confidence intervals ($B_{\text{eff}} = 2000$).

	twist	graft
ann1	0.633 [0.455, 0.787]	0.948 [0.896, 0.987]
ann2	0.780 [0.651, 0.891]	0.966 [0.926, 1.000]
ann3	0.905 [0.791, 0.980]	1.000 [1.000, 1.000]

Table 6: **Intra-annotator agreement.** Bootstrap estimates of Cohen's κ (U-S vs U-SI), clustered by sentence, with 95% confidence intervals ($B_{\text{eff}} = 2000$).

For the Swedish targets, intra-anntoator agreement between paradigms is consistently high. For *konduktör*, κ ranges from 0.762 to 1.000 across annotators. For *motiv*, κ values are even higher, ranging from 0.907 to 1.000. These values indicate substantial to almost perfect agreement, showing that U-S and U-SI produce highly similar annotation outcomes for these words. A similar pattern holds for the English word *graft*, where κ values range from 0.948 to 1.000.

The only notable deviation appears with *twist*, where κ ranges from 0.633 to 0.905. While two annotators show high agreement (≥ 0.78), one annotator exhibits only moderate agreement ($\kappa = 0.633$). Overall, the results do not prove strict equivalence between U-S and U-SI. However, the consistently high κ values for three of the four target words

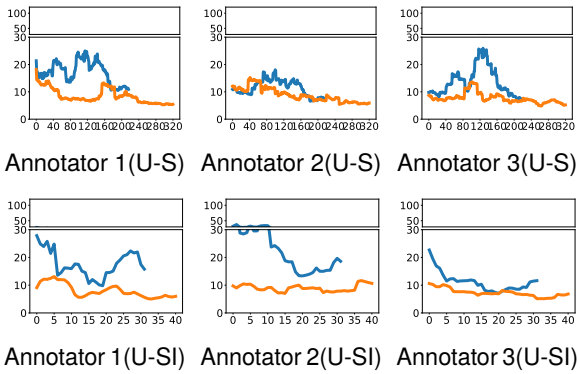

Figure 2: Time in seconds (y-axis) spent by each annotator (shown in the columns) and annotation paradigm (U-S first row, U-SI second row), relative to the number of annotated examples (x-axis) for the words twist and graft.

demonstrate strong practical comparability. Additionally, we note that almost all annotators are more consistent between U-S and U-SI for the second word they annotate, which could be an effect of learning the task. E.g., Swedish annotator 5 has a 0.762 agreement rate for *konduktör* but 0.938 for *motiv*.

Annotation Time To analyze annotator effort over time, we extracted page-level completion times from Qualtrics logs for both paradigms (U-S, U-SI) and both languages (English, Swedish). For each item, the system records the time (in seconds) between page load and submission, which we use as a proxy for annotation duration. We aggregate these times per annotator and per target word, forming sequences of annotation times across items.

Two limitations apply. First, task order was fixed (U-S before U-SI), which may create an order effect if familiarity gained during U-S facilitates U-SI. Second, recorded times likely overestimate true

effort, as pages could be left open before submission. The timings should therefore be interpreted as approximate. To mitigate extreme delays, annotation times were capped at 120 seconds. We then applied moving-average smoothing separately per paradigm: a trailing window of 10 observations for U-SI and 30 for U-S, reflecting longer sequences in the latter. Separate time-course plots were generated per annotator and language (Figures 1 and 2).

Across languages, a consistent pattern emerges after the warm-up phase: U-SI is generally faster and more stable than U-S. Once the sense inventory is internalized, U-SI becomes a direct selection task. In contrast, U-S requires sequential evaluation of each sense, increasing the number of micro-decisions per item.

In Swedish (Figure 1), convergence is relatively smooth, particularly for motiv, where half of the annotators quickly reach low, stable times under U-SI. In English (Figure 2), graft shows similarly efficient behavior, especially under U-SI. By contrast, twist exhibits greater variability and spikes, particularly under U-S, consistent with its lower cross-paradigm agreement.

4.3. Comparison to the U-U paradigm

For each usage, we derive sense labels by aggregating individual annotator judgments through majority voting. For both U-S and U-SI, we count the frequency of each individual sense across annotators and assign the sense with the highest frequency to the usage, applying a random tie-breaking where necessary.

The special label *cannot decide* is not treated as valid assignment in the comparative analysis. We restrict evaluation to usages that receive a valid cluster assignment in all three paradigms (U-U, U-S, U-SI).

Each sense has a corresponding group consisting of all usages that have the sense as a label. To quantify similarity between the diverse usage groupings, we compute the Adjusted Rand Index (ARI) pairwise between U-S, U-S, and U-SI groupings. To assess how the different paradigms capture meaning change, we compute the Graded Change Detection (GCD) score by first deriving, for each method, the distribution of groupings across two time periods and then calculating the Jensen–Shannon divergence between these distributions; higher divergence indicates stronger redistribution of senses across time and is interpreted as greater semantic change.

The results reported in Table 7 show that twist yields the lowest ARI scores across all pairwise comparisons, which mirrors the generally lower agreement observed for this word in all annotation paradigms. In contrast, graft presents a different pattern: while U-S and U-SI are in perfect agreement with each other, their clustering diverges from U-U, resulting in only moderate ARI scores. A close qualitative analysis reveals that in the U-U paradigm one cluster contains a mixture of examples belonging to sense 1 and sense 2, which, although conceptually related through the notion of transplantation, refer to clearly distinct domains, horticulture and medicine. This merging is likely due to annotators frequently assigning an average relatedness score of 3 to such cross domain pairs, which, given the clustering threshold, leads to their aggregation into a single cluster. As a consequence, the semantic change scores differ substantially, with a GCD value of 0.467 in U-U compared to 0.815 in both U-S and U-SI, indicating that U-U underestimates the degree of change for graft by smoothing over a redistribution between domain specific senses across time.

5. Conclusion

In this paper, we examined three annotation paradigms for capturing word meanings over time: Usage–Usage (U–U), Usage–Sense (U–S), and Usage–Sense Inventory (U–SI), with particular focus on the trade-off between scalability, reliability, and expressive power. Through a pilot study on two English and two Swedish target words, we empirically compared these paradigms with respect to inter- and intra- annotator agreement, annotation effort, and their ability to model diachronic semantic change.

Our findings demonstrate that the U–SI paradigm offers a compelling balance between efficiency and quality. While U–U provides high theoretical expressive power by not constraining annotators to predefined sense inventories, in practice this advantage is limited by relatedness labels and the selection of thresholds as well as parameters for the clustering based on heuristics.

Between the two inventory-based paradigms, U–SI emerges as the more efficient option. Cross-paradigm agreement between U–S and U–SI was consistently high for three of the four target words, indicating that the two methods produce largely comparable sense assignments. At the same time, annotation time analyses indicate that U–SI requires less annotator effort once the sense inventory has been internalized, reflecting its single-decision structure compared to the repeated micro-decisions required in U–S. Given that both paradigms scale linearly in the number of usages, the lower cognitive and temporal cost of U–SI makes it preferable for large-scale diachronic annotation projects.

Importantly, **our results do not establish strict methodological equivalence between**

Word	ARI			GCD		
	U-S vs U-U	U-SI vs U-U	U-S vs U-SI	U-U	U-S	U-SI
konduktör	0.703	0.731	0.954	0.603	0.748	0.748
motiv	0.789	0.877	0.895	0.202	0.255	0.307
twist	0.368	0.434	0.717	0.508	0.461	0.445
graft	0.508	0.508	1.000	0.467	0.815	0.815

Table 7: ARI and GCD scores by word and annotation paradigm.

paradigms. Variation across words, most notably for twist, highlights that lexical complexity and sense granularity influence human agreement.

Overall, our pilot study supports replacing usage-pair-based annotation with a usage–sense-inventory framework in large-scale diachronic semantic annotation. U–SI maintains high reliability, reduces annotation time, scales linearly, and preserves meaningful distinctions in semantic change modeling. These properties make it particularly well suited for our long-term objective of annotating word meanings across multiple time periods while operating under realistic budget constraints.

Acknowledgements

This work has in part been funded by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021). We further thank Felix Morger for choosing the senses.

6. Bibliographical References

Eneko Agirre and Aitor Soroa. 2007a. SemEval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 7–12, Prague, Czech Republic. Association for Computational Linguistics.

Eneko Agirre and Aitor Soroa. 2007b. SemEval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 7–12, Prague, Czech Republic. Association for Computational Linguistics.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2004. Correlation clustering. *Machine learning*, 56:89–113.

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

Pierluigi Cassotti, Lucia Siciliani, Lucia Passaro, Maristella Gatto, and Pierpaolo Basile. 2023. WiC-ITA at EVALITA2023: Overview of the EVALITA2023 Word-in-Context for ITAlian Task. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2023)*, Parma, Italy. CEUR.org.

Pierluigi Cassotti and Nina Tahmasebi. 2025. Sense-specific historical word usage generation. *Transactions of the Association for Computational Linguistics*, 13:690–708.

Jing Chen, Emmanuele Chersoni, and Chu-ren Huang. 2022. Lexicon of Changes: Towards the Evaluation of Diachronic Semantic Shift in Chinese. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 113–118, Dublin, Ireland. Association for Computational Linguistics.

Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. 2023a. ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 93–99, Singapore. Association for Computational Linguistics.

Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. 2023b. ChiWUG: A Graph-based Evaluation Dataset for Chinese Lexical Semantic Change Detection. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 93–99, Singapore. Association for Computational Linguistics.

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel word-sense identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1635, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Philip Edmonds and Scott Cotton. 2001a. Senseval-2: Overview. In *Proceedings of The Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*, pages 1–6.

Philip Edmonds and Scott Cotton. 2001b. SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, Toulouse, France. Association for Computational Linguistics.

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

Andrey Kutuzov and Lidia Pivovarova. 2021a. RuShiftEval: A Shared Task on Semantic Shift Detection for Russian. In *Proceedings of the Conference on Computational Linguistics and Intellectual Technologies (Dialogue)*, 20, (online). RSUH.

Andrey Kutuzov and Lidia Pivovarova. 2021b. Three-part Diachronic Semantic Change Dataset for Russian. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 7–13, Online. Association for Computational Linguistics.

Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Enstad, and Alexandra Wittemann. 2022. NorDiaChange: Diachronic Semantic Change Dataset for Norwegian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2563–2572, Marseille, France. European Language Resources Association.

Daniel Loureiro, Aminette D'Souza, Areej Nasser Muhajab, Isabella A. White, Gabriel Wong, Luis Espinosa-Anke, Leonardo Neves, Francesco Barbieri, and Jose Camacho-Collados. 2022. TempoWiC: An evaluation benchmark for detecting meaning shift in social media. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3353–3359, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic. Association for Computational Linguistics.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 240–243. Morgan Kaufmann / MIT Press. SemCor: a WordNet-sense tagged corpus based on the Brown Corpus, manually annotated.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado. Association for Computational Linguistics.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA. Association for Computational Linguistics.

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

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021a. DWUG: A large resource of diachronic word usage graphs in four languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*,

pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021b. DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task in senseval-3. In *Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3)*, pages 41–43.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. LSCDiscovery: A Shared Task on Semantic Change Discovery and Detection in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 149–164, Dublin, Ireland. Association for Computational Linguistics.