Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
D-70569 Stuttgart

Bachelor thesis

Detection of non-recorded word senses

Jonathan Lautenschlager

Studiengang:	B.Sc. Informatik
Prüferin:	apl. Prof. Dr. Sabine Schulte im Walde
Betreuer:	Dr. Dominik Schlechtweg
	Dr. Simon Hengchen
Beginn der Arbeit:	15.06.2023
Ende der Arbeit:	15.12.2023

Erklärung (Statement of Authorship)

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein. [1]

(Jonathan Lautenschlager)

[1]Non-binding translation for convenience: I hereby declare that the work presented in this thesis is entirely my own. I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted hard copies.

Abstract

Wörterbücher enthalten die Bedeutungen von Wörtern zu einem bestimmten Zeitpunkt. Wenn ein Wort in einem Sprachgebrauch eine neue Bedeutung erhält oder eine alte verliert, kann sein Eintrag im Wörterbuch veraltet sein. In dieser Arbeit werden verschiedene Systeme zur Erkennung fehlender Wörterbucheinträge in modernen englischen und schwedischen Wörterbüchern untersucht. Hierfür werden Bedeutungseinträge mit Wortverwendungen aus modernen und historischen Korpora verglichen. Die grundlegende Aufgabe besteht darin zu entscheiden, ob eine Wortverwendung durch einen Eintrag im Wörterbuch abgedeckt ist. Hierfür nutzen wir einen vortrainierten Word-in-Context-Embedder, der es uns erlaubt, diese Aufgabe in einem 'few-shot'-Szenario zu modellieren. Zusätzlich verwenden wir menschliche Annotationen, um unsere Modelle anzupassen und zu evaluieren. Verglichen mit einer zufälligen Probe aus einem Korpus ist unser Modell dazu in der Lage, die Anzahl an nicht abgedeckten Wortverwendungen signifikant zu erhöhen.

Contents

1 Introduction

Dictionaries cover the senses of words at a particular point of time. As they store vast amounts of lexical information in a convenient way, a lot of *Natural Language Processing* (NLP) tasks heavily rely on their quality and completeness. When a word gains a new sense or loses an old one in a speaker community, its dictionary entry may become outdated. Lexicographers regularly check dictionaries for such outdated entries, as pointed out by Lau et al. (2012):

> The manual identification of such new word-senses is a challenge in lexicography over and above identifying new words themselves, and is essential to keeping dictionaries up-to-date. Moreover, lexicons that better reflect contemporary usage could benefit NLP applications that use sense inventories. (Lau et al. (2012))

The aim of the thesis will be to investigate systems that discover missing dictionary entries in modern English and Swedish dictionaries by comparing target word usages from reference corpora to the dictionary entries for the target word. The basic task to solve is to decide whether the sense of a given word usage is covered by any sense in the dictionary entry of the target word or not. The major difficulty is that a dictionary entry, despite containing a good amount of lexical information, does not provide enough context in order to train a standard Word Sense Disambiguation (WSD) system. Depending on the dictionary, the senses of an entry are sometimes only covered by the definition or a few example sentences. We are addressing this problem by comparing different ways of maximizing the quality of information we utilize from the dictionary entries. Further, we make use of Cassotti et al. (2023)'s specialized Word-in-Context (WiC) model XL-LEXEME to create contextualized word embeddings. Although we are not the first to address this problem (Erk, 2006; Lau et al., 2012), none of the related approaches utilizes WiC models to solve the task in a realistic few-shot scenario. Moreover, our model can even be applied to multilingual applications. Furthermore, we use a human annotation approach for a high-quality evaluation of our model's performance.

We start by giving a brief introduction to relevant related work and the underlying concepts of this task in Section 2. We then provide an overview of the various data used in Section 3, followed by a more detailed specification of the task we are attempting to solve in Section 4. The section "Models" (5) describes our different approaches to the task, including all tuning parameters. All experiments carried out are described in detail in the "Experiments" (6) section.

2 Related Work

Detecting non-recorded senses in word usages has, especially with a direct focus on the practical applicability of dictionary maintenance, not been studied extensively (Erk, 2006; Lau et al., 2012). Also known as *Unknown Sense Detection*, the task combines aspects of *Word Sense Disambiguation* and *Word Sense Discrimination*. In the following, we will introduce some of the underlying concepts required for these tasks.

2.1 Task and Models

2.1.1 Word Sense Disambiguation

Word Sense Disambiguation (WSD) is a classical, yet unsolved task in Natural Language Processing (NLP) and has been studied for decades (Weaver, 1955). Navigli (2009) describes the task of WSD as "computationally identifying the meaning of a word by its use in a particular context" (p. 1).

For a more formal approach, we ignore punctuation in a text and refer to the remaining collection of words as a *word sequence* $s = (w_1, w_2, \ldots, w_k)$. A word sequence s that contains a particular target word w, at the position i we call a *word usage* of w and represent it by the tuple $u = (s, i)$. Navigli (2009; p. 4) defines a *word sense* as a commonly accepted meaning of a word. Which meaning is meant depends on the context of use. Let $\Phi = \{\phi_1, \phi_2, \ldots, \phi_k\}$ be a set of predefined senses and $f : U \rightarrow \Phi$ be a mapping from usages to senses given by a lexical resource. The task is to find mapping f.[2] So by definition, WSD can be considered a multi class classification task.

It is also possible to restrict the task so that only senses of the associated target word are considered, i.e., defining the mapping as $f : U \to \Phi_w$, with U being word usages of only a particular target word w and Φ_w being a subset of all senses with only the senses of w.

To correctly classify a word usage based on a predefined set of senses, the sense has to be part of the set in the first place. Building such *sense inventory* with good coverage is strongly dependent on the available data.

Traditional WSD models use large amounts of manually annotated data (multi-shot) to train one *classifier* per word to disambiguate its word usages (Burchardt et al., 2009). Typically, such classifiers have a large number of internal parameters to be optimized (Yarowsky, 1994; Towell and Voorhees, 1998). These training sets are typically a set of word usages associated with their sense label. Since manually labeled data sets are very time-consuming to produce, they are usually of a limited size and do not cover the full lexicon of the language. While these classifier models generally deliver good results on a restricted set of words, they quickly suffer from the problem of data sparseness (Navigli, 2009; p. 4, p. 16). Especially when focusing on certain text types, manually tagged training data becomes very rare. Not only is there fewer data of high quality, but creating it is much more costly as corresponding experts are needed (Navigli, 2009; p. 4, p. 14).

More realistic WSD models therefore try to reduce the dependency on manually labeled training data. The idea is to minimize the number of different parameters the classifier needs to learn, or to share knowledge between different senses. This makes it possible to extract the little training data needed (few-shot) from limited lexical resources like a *dictionary*. A dictionary maps *headwords* to sets of sense entries. A sense entry is usually structured, providing a sense definition (gloss), a number of

[2]The word usages could also be mapped to a collection of possible senses, i.e., an element of the power set of Φ, but since generally only the most appropriate sense is of interest, we directly define it this way.

example usages of the headword in that sense and some meta information such as the etymology.

This is the WordNet data for the headword *unable* in dictionary form:

```
{
    "entries": [
        {
            "pos": "adj.",
            "gloss": "(usually followed by ‘to’) not having the necessary means
                or skill or know-how",
            "examples": [
                "unable to get to town without a car",
                "unable to obtain funds"
            ]
        },
        {
            "pos": "s",
            "gloss": "(usually followed by ‘to’) lacking necessary physical or
                mental ability",
            "examples": [
                "dyslexics are unable to learn to read adequately",
                "the sun was unable to melt enough snow"
            ]
        }
    ],
    "headword": "unable"
},
```

Early few-shot computational approaches to WSD make use of such dictionaries and try to disambiguate word usages by simply counting overlapping words between eligible sense definitions of the target word and sense definitions of nearby words (Lesk, 1986). Newer models often use pre-trained contextualized embedders (Devlin et al., 2019; Peters et al., 2018a) to calculate vector representations for target usages and senses (Section 2.1.4), which can then be compared, e.g. with a similarity metric (Scarlini et al., 2020; Hu et al., 2019; Kumar et al., 2019).

This modelling approach does not require training a parametrized classifier from scratch.

Rachinskiy and Arefyev (2022) develop a model training only one classifier for all words in their data. In this way, knowledge between words can be shared, e.g., from frequent to infrequent ones (Kågebäck and Salomonsson, 2016; Chen et al., 2021).

2.1.2 Word Sense Induction

Word Sense Induction (WSI), also called Word Sense Discrimination, differs from WSD in that it does not assume a predefined sense inventory. It can be thought of as the fully unsupervised counterpart of WSD and aims to group word usages by sense, instead of classifying word usages (Schütze, 1998; Erk, 2006).

When referring to WSD as a multiclass classification task, WSI can be seen as a clustering task. Hence, we also formalize it as such (Diday and Simon, 1976; p. 48): Let $U = \{u_1, u_2, \ldots u_n\}$ be a set of word usages of a word w. Let $\Psi = \{\psi_1, \psi_2 \ldots \psi_k\}$, $k \leq |U|$ be a set of labels representing non-structured senses for w and a mapping $f : U \rightarrow \Psi$ given by a lexical resource. The elements of a resulting equivalence class C_ψ can now be seen as different usages of a word w with the same sense ψ.

$$C_\psi = \{u | f(u) = \psi\}$$

The task is to find the set of equivalence classes $C = \{C_\psi | \psi \in \Psi\}$. WSI has applications in lexicography (Lau et al., 2014) and lexical semantic change detection (Lau et al., 2012; Laicher et al., 2021; Martinc et al., 2020). Moreover, it can be used to improve the results for highly ambiguous queries in web information retrieval (Véronis, 2004), in web search clustering to help search engines deliver more diverse results (Navigli and Crisafulli, 2010) and in the enrichment of lexical resources such as WordNet (Miller, 1994; Nasiruddin et al., 2014).

2.1.3 Unknown Sense Detection

The task of Unknown Sense Detection (USD), as defined by Erk (2006), combines aspects of WSD and WSI. While there are predefined senses, the task is not to

	Task	Sense Inventory	Supervised
WSD	Multiclass classification	predefined	yes
WSI	Clustering	none	no
USD	Binary classification	predefined	yes

Table 1: Differences between WSD, WSI and USD

map word usages to these, but instead to find word usages whose meanings are not covered by the sense inventory. It is of no further importance by which known sense (if any) a word usage is described (Erk, 2006). Therefore, it can be seen as a binary classification task: Let $U = \{u_1, u_2, \dots u_n\}$ be a set of word usages of a word w. Let $\Phi = \{\phi_1, \phi_2, \dots, \phi_k\}$ be a set of predefined senses and $f : U \to \Phi$ be a mapping from usages to senses given by a lexical resource. Let further $g : U \to \{0, 1\}$ be a mapping such that $g(u) = 1$ iff f is undefined and $g(u) = 0$ otherwise. The task is to find mapping g. That is, word usages that are not covered by any entry in our sense inventory need to be assigned label 1 while covered usages need to be assigned label 0.

One of the few modeling approaches to USD is given by Erk (2006), measuring the distances between the vector representations of target usages and sense entries. If the target usage deviates far from all sense entries, i.e., for all sense entries another sense entry is closer than the target usage, it is considered an outlier (Erk, 2006).

Although all tasks share fundamental characteristics, there are significant differences between them. Table 1 presents an overview of the tasks.

2.1.4 Contextualized Vector Representations

So-called *contextualized embeddings* are numeric, high-dimensional vector representations of word meaning. These are usually learned as parameters in a *language model* trained on large amounts of data (Devlin et al., 2019; Peters et al., 2018b). They encode distributional information (Harris, 1954; Firth, 1957) for words in context and can be used to measure the semantic similarity between word usages (cf.

Schütze, 1998).

BERT is one such contextualized embedding model that generates a high-dimensional vector representation for a given context. The model architecture is based on the original multi-layer bidirectional Transformer encoder by Vaswani et al. (2017). BERT-embeddings find use in various NLP tasks (Devlin et al., 2019). Sentence-BERT (SBERT) extends BERT, using siamese and triplet networks to produce meaningful sense embeddings for a single sentence (Reimers and Gurevych, 2019). XL-LEXEME is a pretrained contextualized embeddings model that is based on SBERT and fine-tuned on human-labeled Word-in-Context (WiC) (Pilehvar and Camacho-Collados, 2019) data.[3], to study lexical semantic change. Its structure gives prominence to a target word in the given context.

3 Data Sets

3.1 Corpora

Since our goal is to support dictionary maintenance, we have to make sure to create a realistic scenario. Use of language changes over time, which necessarily leads to the emergence of sense entry gaps in dictionaries. Non-recorded word senses can occur for two main reasons: Either they are old senses that were left unrecorded when the dictionary was created, or they are novel senses that emerge due to language change. The former are likely to be found in historical data, while the novel senses more likely occur in modern corpora. To train our models on both cases, we decided to use modern and historical corpora.

Table 2 shows an overview of our corpora. A precise statistical description of the modern corpora can be found on the website of Wortschatz Leipzig.[4]

[3] Decide for two given word usages, whether they have the same sense or not (Pilehvar and Camacho-Collados, 2019)

[4] https://cls.corpora.uni-leipzig.de/

	Modern		Historical	
Language	**English**	**Swedish**	**English**	**Swedish**
Name	Leipzig_News	Leipzig_News	CCOHA	Kubhist2
Year	2020	2022	1810–1860	1790–1830
Source	Goldhahn et al. (2012)		Alatrash et al. (2020)	Språkbanken (2019)
Sentences	1 million	1 million	~250 thousand	3.3 million

Table 2: Comparison of corpora

3.1.1 Modern

We decided to use the latest *News* datasets from the Leipzig Corpora Collection (Goldhahn et al., 2012) for both English and Swedish. These datasets contain each 1 million sentences in randomized order retrieved from articles on news websites published in the years 2020 (English) and 2022 (Swedish), without further processing. Due to the great similarity in origin and nature between these two corpora, we already achieve a solid basis for the comparability of both languages.

3.1.2 Historical

For Swedish, we use the first corpus of the Swedish test data for the SemEval 2020 Task 1 (Schlechtweg et al., 2020). It is a sample of the KubHist2 corpus (Språkbanken, 2019) containing sentences in randomized order from the period 1790–1830. The corpus consists of about 71 million tokens, but contains frequent OCR errors (Adesam et al., 2019).

For English, we used a corpus based on CCOHA (Alatrash et al., 2020), a cleaned version of the well-known diachronic COHA corpus of American English (Davies, 2012). This corpus contains sentences in randomized order from the period 1810–1860 and consists of roughly 6 million tokens.

3.2 Lexical Resources

3.2.1 English

WordNet 3.0 (Miller, 1994; Fellbaum, 2005) is a well-known and well-established large lexical database for the English language. It groups words which are synonymous in one of their meanings into so-called *synsets* that describe distinct concepts or senses. Each synset has one *primary headword*, under which the synset can be listed in a dictionary-like structure. We refer to these as *primary headwords*. The word `car`, for example, has five senses assigned, i.e., is part of five synsets, including the following two:

<**car**, auto, automobile, machine, motorcar>

<cable car, **car**>

The first synset, defined by its gloss as "a motor vehicle with four wheels; usually propelled by an internal combustion engine", has `car` as the primary headword, followed by four other headwords from WordNet. It also includes the example usage "he needs a car to get to work". The second synset, defined by its gloss as "a conveyance for passengers or freight on a cable railway" has `cable car` as the primary headword, while `car` is only an additional headword. It includes the example usage "they took a cable car to the top of the mountain".
There are a total of 117,000 synsets, all containing a gloss. Only a portion also includes example usages that put a member of the synset into a usage context with the respective meaning.

There are two ways to restructure WordNet to transform it into a dictionary-like structure, i.e., subordinating senses to headwords: Either assign synsets only to the primary headword and accept definition gaps in return, or assign synsets to all participating headwords, but end up with duplicate senses in the dictionary. We have opted for a mixture of both, by taking all synsets into account, but keeping track of whether it is a synset that has the headword as primary headword or not. If this is the case, we refer to them as *primary synsets*. In the following, we explicitly specify when only primary synsets are meant.

3.2.2 Swedish

As part of this project, we have been granted access to Svensk Ordbok (SO)[5], a dictionary of the Swedish language, created and maintained by the University of Gothenburg (Allén, 1981). The data is private, but there is an online interface for the dictionary. It is structured in the classic style of a dictionary in that it lists senses under a headword. In order to use a standardized form for such a headword, Allén (1981) apply the *lemma-lexeme model* and groups different inflections and forms of words by their lemma. In this way, SO stores a total of 68,000 senses for over 41,500 headwords. The majority of the senses are described by a gloss, in some cases extended by a *secondary gloss*.[6] All sense entries contain example usages which serve as examples of the use of the headword in the respective sense.

3.2.3 Comparison

The two dictionaries[7] represent the senses of words, describe them using both gloss and example usages, and group them by their headword.

Both dictionaries differ in that their entries have gaps in different parts of their sense entries. While WordNet senses are fully covered by gloss but only have partial example usages, it is the other way around in the Swedish dictionary. A direct comparison of the statistics of both dictionaries is shown in Table 3.

4 Task

For a set of given word usages $U = \{u_1, u_2, \ldots, u_k\}$ of a word w, we want to solve the binary classification task of USD as defined in Section 2.1.3, to detect non-recorded senses in the corresponding sense inventory. Let $\Phi = \{\phi_1, \phi_2, \ldots, \phi_n\}$ be the sense

[5] `https://svenska.se/so/`

[6] We only made use of the secondary gloss if no standard gloss was given.

[7] We use the term "dictionary" for WordNet here, after transforming it into a dictionary-like structure (see Section 3.2.1).

	WordNet	Swedish Dictionary
Headwords	86,555	41,597
Avg. senses per headword	1.36	1.64
Avg. senses per headword with multiple senses	2.27	2.91
Percentage of senses with gloss	100%	79%
Avg. length of gloss	56.40	34.28
Percentage of senses with examples	28%	100%
Avg. number of examples per sense	0.41	3.37
Avg. examples per sense with examples	2.84	3.37
Avg. length of examples	33.60	32.44

Table 3: Direct comparison of both dictionaries

inventory, i.e., a set of predefined senses, obtained from a dictionary. Let $f : U \to \Phi$ be a mapping from usages to senses and $g : U \to \{0, 1\}$ be a mapping such that $g(u) = 1$ iff f is undefined and $g(u) = 0$ otherwise. Our task is to find mapping g, i.e., assigning the label 1 to word usages that are not covered by our sense inventory.

5 Models

All our models have the same basic structure but differ in the type of data used and the embedding comparison methodology. Essentially, they first use the information given in the dictionary to create a vector representation for each existing sense. The type of information used (e.g., gloss, example usage) and how it is processed (e.g., inserting headwords if missing) are treated as model hyperparameters, as listed in Section 5.1. The models then create a usage embedding for the word usage to be examined and calculate the similarity to all sense embeddings of the associated headword. If the calculated similarity score between the usage embedding and the closest sense embedding is below a specific threshold, the model predicts that the word usage is the occurrence of an unknown sense and therefore marks it as unassigned.

We test and compare every possible combination of settings for the following hyperparameters, as far as the available data allows.

5.1 Hyperparameters

Each model's identifier is made up of the different choice of the hyperparameters (see the following sections for details):

- sense embedding $\in$ `[G0, G1, G2, G3, E0, E1, E2, E3, E4]`,
- usage embedding $\in$ `[`ϵ`, SUB]`,
- similarity measure $\in$ `[COS, SPR]`,
- threshold $\in \{0.0, 0.01, 0.02, \ldots, 1.0\}$

An example would be `G3_SUB_COS_0.48`, for the model using the gloss of sense entries after applying replacement strategy 3, substituting the target word with the headword in the usage embedding, evaluating the similarity between embeddings by using *Cosine Similarity* (Salton and McGill, 1983) and deciding between assigned and unassigned word usages based on a threshold of 0.48.

5.1.1 Sense Embedding

We use the XL-LEXEME model (see Section 2.1.4) to create contextualized embeddings that represent the senses in our sense inventory. We decided to use XL-LEXEME because of its good performance in the LSCD task, as the ability to extract enough information from a single word usage to examine lexical semantic change provides an excellent basis for our setting. The model encloses the target word in the word usage with special symbols. However, glosses in dictionaries typically do not include the actual headword itself and therefore do not provide a target position. A similar problem can occur with example usages: In WordNet, the example usages of a synset always contain one headword from its synset, not necessarily the primary headword. Therefore, not even the example word usages are suitable

option	description	example
0	leaves the sequence as is	a poor salary
1	HEADWORD: SEQUENCE	inadequate: a poor salary
2	SEQUENCE (HEADWORD)	a poor salary (inadequate)
3	SEQUENCE, i.e., HEADWORD[8]	a poor salary, i.e., inadequate
4	replace word[9]	a inadequate salary

Table 4: Replacement strategies on an example of *inadequate* taken from WordNet

without restriction. Using various strategies, we therefore modify the glosses and example word usages if necessary. In Table 4 all strategies are explained using an example usage taken from WordNet for a *primary synset* of the headword *inadequate*. Instead of the headword itself, the example usage contains the headword *poor*, which is also part of the synset.

By applying these replacement strategies on both example usages and gloss, we receive the gloss models `G[0-3]` and the example models `E[0-4]` and `E[0-3]` for English and Swedish respectively. The former represent a sense only by its gloss while the latter represent it only by its example usages. In the case of multiple example usages we take their average embedding. Note that replacement strategy 4 is only used on English example usages as this is the only data where synset replacement is possible and necessary.

The models are naturally limited in their predictions by gaps in the data, i.e., missing gloss or example usages. If a headword has no glosses or no example usages for all its senses, it is not represented by the sense inventory of the corresponding model and can therefore not be examined. Headwords where individual senses are not covered by the data are part of the sense inventory. However, the predictions are only of limited practical relevance, since word usages of senses that are actually represented but have incomplete entries are predicted as unassigned.

[8]"dvs." instead of "i.e." for the Swedish version.

[9]If there is no headword of the synset present, strategy 2 is used.

5.1.2 Target Usage Embedding

XL-LEXEME is used to generate a contextualized embedding from the entire target usage and the position of the target word. The usages are obtained by directly searching for variations of the headwords (Section 6.2.1), thus a headword must be contained in the usage in some form. Technically-speaking, therefore, no replacement strategy is necessary to create an embedding. However, we still test a variation in which the target word is replaced by the headword (often lemma), through the models `G[0-3]_SUB and E[0-4]_SUB`, inspired by the results of Laicher et al. (2021).

5.2 Embedding Comparison

The similarity between the target embedding and each sense embedding is calculated using both Cosine Similarity and Spearman's rank correlation coefficient (Spearman, 1904). As Spearman shows better performance than Cosine Similarity for the WiC task in (Tabasi et al., 2022).

5.2.1 Threshold

A later fine-tuned threshold decides between *assigned* and *unassigned* word usages based on the comparison scores. We test values in the range of $\{0.0, 0.01, 0.02, \ldots, 1.0\}$.

6 Experiments

Our experiments can be roughly divided into two parts:[10] The first consists of an initial phase of human annotation on a random sample from the corpora. The obtained data from this phase is used to tune the hyperparameters and to estimate the performance of the different models. In the second part, the most promising models are used to find unassigned usages in subsets of the corpora. The human annotation is then repeated on the found usages to determine the performance of the models.

[10] All code and data for the experiments can be found on GitHub: `https://github.com/joni0700/non-recorded-sense-detection`

6.1 Annotation

As part of this project, we carry out two phases of human annotation. The first phase is conducted both in order to collect data for the tuning of our models and to establish a reference point for the second annotation. The second phase serves the evaluation of the quality of our models' predictions.

The annotators are presented a word usage with the target word marked and a sense gloss that may or may not describe this usage. Additionally, they are shown the list of all possible sense glosses, to better assess the presented one (see below). In our annotation, we take inspiration from Erk et al. (2013)'s WSsim method by conducting individual assessment of all senses for a usage, but differ in that we ask for a binary classification into *sense gloss fits* (label "1") and *sense description does not fit* (label "0"), instead of a five-level rating. The annotators also always have the option to choose that no specification is possible (label "-"). They are encouraged to leave a comment if this is the case.

Find an example of an annotation instance for a usage of the word *relative* in Figure 1. The word itself is highlighted, one gloss is marked for annotation and all glosses are given as additional information.

A total of six annotators, three for each language, are recruited. All annotators are students and native speakers of the respective language. Before the annotation, they received a 30-minute briefing and underwent a short test annotation to familiarize with the process. Both rounds of annotation were carried out using the PhiTag platform.[11] The data for the test annotation comes from the annotation_standardization repository.[12][13]

[11] https://phitag.ims.uni-stuttgart.de/

[12] https://github.com/ChangeIsKey/annotation_standardization/tree/main/use_single/wsbest/english/tutorial

[13] We provide annotation guidelines, tutorial and annotated data at https://github.com/joni0700/non-recorded-sense-detection

```
usage:
Abigail's sister accompanied her to this meeting because
the sister alleged that she had witnessed some of the childhood
sexual abuse from the relative.

gloss:
estimated by comparison; not absolute or complete

all glosses:
estimated by comparison; not absolute or complete

an animal or plant that bears a relationship to another
(as related by common descent or by membership in the same genus)

a person related by blood or marriage
```

Figure 1: Sketch of an annotation instance for a usage of the word *relative*.

6.2 Part 1

6.2.1 Annotation

Human annotation is conducted on a random sample from the corpora. The annotations were carried out identically for both languages, in terms of data sampling and briefing the annotators. Word usages from the modern and historical corpora were retrieved by lemmatizing the sentences and searching for appearances of a randomly selected subset of headwords from the language's dictionary.

In each corpus, we search word usages of a random sample of 3000[14] headwords until we find at least one usage for 150 of them. We keep at most five usages for each headword, chosen at random. Lastly, we combine the word usages from the modern

[14]We use a sample of 3000 headwords because comparing every lemma in every sentence with all headwords in the dictionaries would be too much computing effort.

and the historical samples. This results in a set of word usages that is approximately equally distributed between modern and historical corpora, which can be linked to a headword in the dictionary. Combining each word usage with each eligible sense returns approximately 1,200 annotation instances, distributed over 500–700 usages, as listed in Table 5. For WordNet, only primary sysnsets are considered in the first annotation phase. These are assessed by the three annotators. After the first annotation phase, one English annotator was excluded for judgement inconsistencies and a new annotator was recruited.

After the annotation, we aggregate judgements per annotation instance by majority. Then, we aggregate these majority labels by usage to decide whether the usage is assigned or unassigned: A usage is considered assigned iff at least one instance has the majority label “1”.

6.2.2 Annotation Analysis

Find a summary of the annotation results in Table 5 and Table 6. For English, out of 473 usages, 45 are labeled unassigned (9.5%), i.e., all instances have the majority label “0”. For Swedish, out of 674 usages, 95 are labeled unassigned (16.6%). The inter-annotator agreement measured by Krippendorf’s alpha (Krippendorff, 2018) as shown in Table 6 is low to moderate. Especially in English, there is considerable disagreement. The pairwise comparison reveals, that annotator A1 stands out in particular. This applies in the same way to both, modern and historical usages.[15] For Swedish, the annotators are more in agreement, especially on the modern data.

6.2.3 Model Selection

In order to not overfit the models on the limited data, we assess them using k-fold cross-validation. We perform a total of 10 rounds of 5-fold cross-validation for every model. Each round consists of different manipulations of the gold sense assignments. In each round, we randomly mask a set of assigned senses as unassigned in order to

[15]This annotator was excluded from the second annotation phase.

	English	Swedish
Instances	1165	1202
Usages	474	706
Label distribution (0, 1, -)	(1840, 1651, 4)	(1294, 2104, 208)
Excluded instances	2	87
Remaining usages	473	674
Assigned	428	562
Unassigned	45 (9.5%)	95 (16.6%)

Table 5: Statistics of the first human annotation phase

	English			**Swedish**		
	All	**Modern**	**Historic**	**All**	**Modern**	**Historic**
A1 vs. A2	0.331	0.392	0.276	0.408	0.506	0.317
A1 vs. A3	0.273	0.333	0.220	0.575	0.614	0.527
A2 vs. A3	0.561	0.580	0.544	0.524	0.655	0.353
Full	0.389	0.434	0.349	0.480	0.588	0.371

Table 6: Inter-annotator agreement using Krippendorff's alpha of annotation phase 1. The labels A1–A3 represent different annotators across languages. The label "-" is excluded from the data before calculating agreement

simulate unknown senses (cf. Erk, 2006).[16] The models have to know at least one sense of a headword to examine its usages (see Section 5. Therefore, masking can under certain conditions lead to loss of data. In order to prevent this, we have to consider three special cases:

1. Senses of headwords that have no complete sense are excluded from the model evaluation since its word usages cannot be compared against any sense vector.

2. Incomplete senses are always masked as a model could never assign this sense.

[16] Unassigned senses are left untouched, but are still included in the evaluation data.

3. Senses that are the only complete sense of the headword are always unmasked as masking them would exclude it from evaluation.

Note that headwords of all remaining senses have multiple complete senses and can therefore be randomly masked.

We use the same simulated data for all models of a language, even though the gloss models in English and the example models in Swedish do not suffer from limited dictionary completeness. To further ensure the model's capability to examine every word usage after masking, we randomly select one sense for each headword to exclude before masking all remaining. This prevents us from accidentally masking all senses of a headword.

Based on this random masking, we label all word usages from the human annotation *assigned* if at least one assigned sense is not masked. Assigned represents the negative label in our evaluation while unassigned represents the positive labels. All other word usages are labeled *unassigned*. This set of word usages is then randomly divided into 5 folds for the 5-fold cross-validation.

Applying the randomization ten times leaves us with ten different data sets of fivefold subdivided usages. In each round for each fold k, the model uses the data of the remaining four *training*-folds to determine a threshold $\in \{0.0, 0.01, 0.02, \ldots, 1.0\}$ (see Section 5.2.1) for the similarity measure, so that a F_β-score is maximized for $\beta\ in \{0.1, 0.3, 0.5\}$ separately to check model predictions with varying importance of precision. Using this threshold, the model predicts on the *test*-fold k. The F_β-score modifies the more common F-score so that a recall factor β changes the importance of the recall. By choosing a $\beta < 1$, precision is rated more important. The smaller the β the more important becomes the precision. Precision is the share of true positives in the instances predicted by the model as positive. Hence, optimizing for precision will increase the number of true positives in our predictions. Prioritizing precision will also tend to decrease the number of usages labeled by the model as assigned because choosing the upper percentiles of a (meaningful) rank-cutoff-based classifier will tend to decrease the probability for a false positive, as apposed to including

	Average		Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test
Threshold	0.478		0.480		0.480		0.480		0.470		0.480	
Precision	0.726	0.691	0.643	1.000	0.690	0.857	0.750	0.500	0.741	0.600	0.808	0.500
Recall	0.390	0.378	0.383	0.421	0.357	0.600	0.436	0.182	0.400	0.188	0.375	0.500
$F_{0.5}$	0.618	**0.572**	0.566	**0.784**	0.581	**0.789**	0.656	**0.370**	0.633	**0.417**	0.656	**0.500**
random_ $F_{0.5}$	0.356	0.262	0.333	0.360	0.359	0.081	0.350	0.182	0.364	0.341	0.375	0.345
frequency_ $F_{0.5}$	0.534	0.446	0.485	0.571	0.481	0.741	0.546	0.263	0.563	0.333	0.593	0.321

Table 7: Cross-validation results in round 10 of model `E4_COS` when maximizing $F_{0.5}$-score on English data

lower percentiles.[17] In terms of practical application, we deem a small yet precise sample as much more useful than a large, imprecise one.

In addition, we establish baselines to be able to better estimate the performance. This includes a random-baseline that predicts *assigned* with a probability of p where p is equal to the share of assigned usages in the gold data, and a frequency-baseline that only predicts assigned if usage can be assigned to its most frequent sense and unassigned otherwise. This baseline only applies to WordNet where we have sense frequency information.

Find an example for the results of one round of cross-validation in Table 7. We calculate the average of all values for the test fold and the training folds for each round. Figure 2 shows the precision-recall vs. threshold curve and nicely illustrates the differences between the different scores: While $F_{0.1}$ accepts almost any recall in order to maximize the precision, $F_{0.3}$ and $F_{0.5}$ find a balance between the two values.

After manually checking model predictions and analyzing performances we chose $F_{0.3}$-score for model selection. While $F_{0.1}$-score would be even better in terms of precision, we observe in the individual rounds of the cross-validation that a β of 0.1 is too extreme and excludes a disproportionate number of true positives. When optimizing for $F_{0.3}$-scores, model `E4_COS` produces the best results for English, as shown

[17] We rank usage-sense pairs by similarity (see Section 5.2) and treat them as assigned based on the threshold parameter (unassigned below, assigned above), which will determine the class label of the usage. Hence, the lower the cutoff, the larger the sample. And because the probability for including a true positive is expected to increase with a higher cutoff, a lower cutoff will likely decrease the precision.

$F_{0.3}$-score								
	English				**Swedish**			
	DEFAULT		**SUBST.**		**DEFAULT**		**SUBST.**	
	COS	**SPR**	**COS**	**SPR**	**COS**	**SPR**	**COS**	**SPR**
E0	0.466	0.523	0.477	0.452	0.417	0.419	0.413	0.400
E1	0.589	0.590	0.583	0.592	0.451	0.443	0.413	0.399
E2	0.579	0.562	0.566	0.590	0.425	0.431	0.411	0.408
E3	0.494	0.523	0.493	0.489	0.428	0.431	0.397	0.392
E4	**0.613**	0.584	0.593	**0.612**				
G0	0.270	0.280	0.255	0.267	0.349	0.371	0.345	0.340
G1	0.227	0.220	0.226	0.209	0.600	0.606	0.549	0.537
G2	0.260	0.223	0.264	0.245	0.550	0.612	0.564	0.547
G3	0.217	0.259	0.275	0.256	**0.625**	0.617	0.599	0.621

Table 8: Results of cross-validation. Performance is given as average $F_{0.3}$ across rounds and folds

in Table 8. However, since the value deviates only minimally from the performance of `E4_SUB_SPR`, we decided to use the latter, as the standard deviation of the scores over ten rounds of cross-validation is lower. For Swedish, we decided on `G3_COS`, as it achieved the best results.

6.3 Part 2

6.3.1 Annotation

The two models predict on equally-sized subsets of 100–150 thousand sentences, for both modern and historical corpora, respectively. The sentences are filtered by excluding those that are longer than 300 characters or have too many punctuation tokens (> 25%). Every sentence is lemmatized and then searched for the lemmas of the headwords represented in the respective model's sense inventory. When a headword is found, the model predicts for this usage, whether it is assigned or

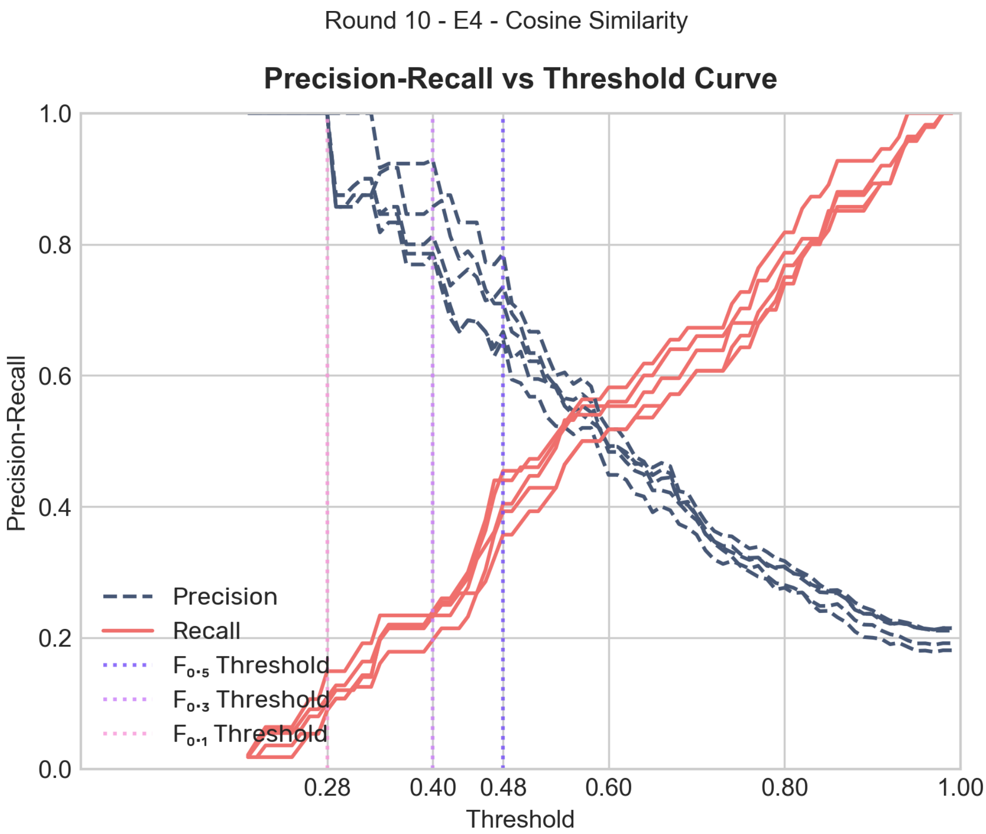

Figure 2: Precisions and recalls of all five folds in the cross-validation round 10 of model `E4_COS` on English data

not. The resulting prediction sample consists of 3608 usages and 706 headwords in English and 12534 usages and 2061 headwords in Swedish.

In a second step, we further process this sample in the following ways:

- We exclude partially complete headwords (at least one sense incomplete).
- Unassigned usages are sorted by similarity to the nearest sense, i.e., the usage that is the furthest from all eligible senses is first.

Then, usages for evaluation are chosen from the top of this list. At most eight usages for the same headword are sampled to ensure a broad analysis. These usages were combined with all eligible senses of the according dictionary. For WordNet, this time not only primary synsets are considered (in contrast to phase 1) in order to prevent missing senses recorded under another headword. (We consider this difference to the first round in the evaluation below.)

The finally resulting sample has roughly the same size as in the first annotation phase. This is uploaded to PhiTag and annotated in parallel to the first phase. Find more detailed statistics in Table 9.

6.3.2 Annotation Analysis

For English, after exchanging one annotator, the annotator-agreement on the full data is 0.384, as displayed in Table 10. The agreement is comparable to the first phase, however, note that we increased the number of senses the annotators can choose from by including the number of non-primary synsets. There is considerably higher consensus among the Swedish annotators (0.56). Table 10 shows that this is also the case for the pairwise comparison. For English, agreement on historical instances is slightly higher than on the modern instances. This contrasts with phase 1. For Swedish, it is considerably worse on historical instances, as in phase 1.

6.3.3 Model Evaluation

The second (prediction) phase yields surprisingly diverse results. While both models successfully increase the number of unassigned usages in the sample data, the Swedish model is considerably more successful. Table 9 shows the results. Out of 1000 predictions, the Swedish model is correct in almost two thirds of cases according to human judgment. In the predictions of the Swedish model, there is a significantly higher number of usages with few senses (1400 instances for 1001 usages vs. 1208 instances for 322 usages, even though we have similar numbers of senses per headword for both dictionaries, as shown in Table 3).

Note that in English, the samples from both phases are not completely comparable because secondary synsets were excluded in the first phase. However, taking them into account for the first phase would only reduce the number of unassigned

	English	Swedish
Instances	1208	1400
Usages	322	1001
Label distribution (0, 1, -)	(2151, 1462, 11)	(2529, 1218, 456)
Excluded instances	5	109
Remaining usages	322	927
Assigned	277	327
Unassigned	45 (13.98%)	600 (64.725%)

Table 9: Statistics of the second human annotation phase

usages, which would only strengthen our case. In order to make the two samples more comparable, we post-hoc exclude secondary synsets, recalculate the results and observe that...

Differentiating between modern and historical instances indicates a commonality between both languages: Table 11 and Table 12 display the results of both phases of human annotation for English and Swedish, respectively. In the first phase, unassigned senses were almost equally distributed between modern and historical usages in the English data. In the second phase, however, the proportion of correctly predicted historical usages hardly differs from the random sample of the first phase while the correct predictions of modern usages increases by two thirds. We observe a similar trend in Swedish: Although the distribution of unassigned usages is very different between modern and historical in the first Swedish phase, with proportionately almost twice as many unassigned usages in the historical data, a stronger increase in the modern usages can also be observed here.

6.3.4 Manual Analysis

We perform a manual analysis of the true positives from the English model predictions as the authors are speakers of that language, in contrast to Swedish. The figures below show some of the true positively predicted usages.

	English			Swedish		
	All	**Modern**	**Historic**	**All**	**Modern**	**Historic**
A1 vs. A2	0.335	0.310	0.347	0.539	0.566	0.456
A1 vs. A3	0.342	0.291	0.397	0.558	0.546	0.552
A2 vs. A3	0.464	0.490	0.374	0.608	0.641	0.526
Full	0.384	0.368	0.375	0.559	0.584	0.490

Table 10: Inter-annotator agreement using Krippendorff's alpha of annotation phase 2. The labels A1–A3 represent different annotators across languages. The label "-" is excluded from the data before calculating agreement

	Phase 1			Evaluation Phase		
Usages	**All**	**Modern**	**Historical**	**All**	**Modern**	**Historical**
total	474	232	242	322	210	112
excluded	1	1	0	0	0	0
remaining	473	231	242	322	210	112
assigned	428	208	220	277	176	101
unassigned	45 (9.5%)	23 (9.9%)	22 (9.1%)	45 (13.98%)	34 (16.2%)	11 (9.8%)

Table 11: Comparison of the different corpora of the English data

We start with cases where also a close manual analysis suggests that they are truly non-recorded in our dictionary. Consider the following word usage of *pipeline*:

usage

There should be some things that can be done in the short term, but in terms of developing the pipeline further on coaching and executive positions, that would take a longer period of time

senses

gossip spread by spoken communication

a pipe used to transport liquids or gases

Usages	All	Phase 1 Modern	Historical	All	Evaluation Phase Modern	Historical
total	706	337	369	1001	478	523
excluded	52	4	28	74	9	65
remaining	674	333	341	927	469	458
assigned	562	293	269	327	224	103
unassigned	112 (16.6%)	40 (12.0%)	72 (21.1%)	600 (64.7%)	245 (52.2%)	355 (77.5%)

Table 12: Comparison of the different corpora of the Swedish data

This metaphoric use of the word is not recorded in WordNet. Equally, for the following usage of *blacken*, *baked* and *qualification*:

usage

and especially ministers , leave a church from base motives , they of course and proverbially hate their former friends , depreciate their talents , and blacken their names .

senses

make or become black

burn slightly and superficially so as to affect color

usage

No wonder hes up there getting big baked.

senses

cook and make edible by putting in a hot oven

prepare with dry heat in an oven

heat by a natural force

be very hot, due to hot weather or exposure to the sun dried out by heat or excessive exposure to sunlight

(bread and pastries) cooked by dry heat (as in an oven)

usage

Slovan Bratislava's Champions League qualification tie against Faroe Islands side KI Klaksvik has been cancelled after a player from the Slovakian club tested positive for COVID19, European soccer's governing body UEFA said on Saturday.

senses

an attribute that must be met or complied with and that fits a person for something

the act of modifying or changing the strength of some idea

a statement that limits or restricts some claim

We observe various historical usages with likely historical meanings that are seemingly non-recorded:

usage

You seem to intend a eulogy , yet leave out whatever was noblest in her , and blacken while you mean to praise .

senses

make or become black

burn slightly and superficially so as to affect color

Note that we found two distinct non-recorded word senses for *blacken*. However, true positives also contain a number of problematic cases. A major problem are multi-word expressions. Our word usages sampling algorithm (see Section 6) has only very limited capability to detect multi-word expressions like *revolves around* in the following example:

usage

Im at that age where many of my friends are having children, and a central topic of conversation whenever were together revolves around creating the almostscientifically set schedule for their babies.

senses

turn on or around an axis or a center

move in an orbit

cause to move by turning over or in a circular manner of as if on an axis

Our sampling algorithm only detects the headword *revolves* and thus compares it to the wrong sense entries. Had it detected *revolves around*, it could have assigned the usage to the correct sense existing under the headword *revolve_ around*. While this kind of error is frequent, the prediction is technically correct. Note, however, that there are cases where multi-word expressions, like idioms, are involved, but no sense entry is present: *to carry the flame of reforms*, or *to go at length*. There are also cases where an ellipsis of a represented multi-word occurs. For these, one could argue that they should receive a sense entry in the sub-headword. Consider the following example: While WordNet has the headword *wiped out* with a sense gloss matching the usage, the usage only contains *wiped*, for which the model predicts all senses unassigned:

usage

Notably, the local Native American tribes were not targeted or wiped the new nation embraced them as equals and allowed the tribes a major say in the rule of America.

senses

rub with a circular motion

Furthermore, OCR errors in the data can also lead to technically correct *unassigned* predictions while there is a matching sense entry under a different headword. It is very likely that the following example is a misspelling of the headword *shark*:

usage

to the river and washed , first for , it was in vain to think of swimming there , at a season of the year , when the fresh waters , all about , were full of sawdust , slabs , mud turtles , and floating timber to say nothing of water snakes , which were far more terrible to me than **shirks** .

senses

avoid (one's assigned duties)

avoid dealing with

Incorrect lemmatization can have a similar effect like in the following example of *virtually*:

usage

It will take time to adjust to all that is needed from building routines, learning inperson or **virtually**, teaching in both venues simultaneously, and striving to be together when we are physically apart.

senses

in essence or effect but not in fact

(of actions or states) slightly short of or not quite accomplished; all but

The adverbial usage of *virtual* is incorrectly mapped to the lemma *virtually* in this example.

We also observe various likely ambiguous usages:

usage

As previously stated in multiple letters from the installation to MDOT, the Navy will not cede any property for the construction of this toll road, Seymour writes.

senses

the act of installing something (as equipment)

a building or place that provides a particular service or is used for a particular industry

a formal entry into an organization or position or office

usage

Stolen bases are a thing for me, Betts said.

senses

a relation that provides the foundation for something

the fundamental assumptions from which something is begun or developed or calculated or explained

the most important or necessary part of something

We conjecture that both model and annotators may tend to annotate ambiguous usages as unassigned.

We also find some ad-hoc innovative word meanings, which are likely not lexicalized, like the following use of *flame*:

"But sloppy, agendadriven journalism of this sort fans the flames of racial division."

7 Conclusion

Our goal is to automatically detect non-recorded word senses in a large historical and modern corpora, based on a given large, but possibly imperfect sense inventory. Our model uses pre-trained WiC models to generate sense embeddings from the little data a dictionary provides, compares them to the contextualized embedding of a word usage, created by the same model and decides, based on a threshold for similarity, whether any sense from the inventory is expressed by the usage.

Our method considerably increases the chance to find non-recorded word senses in corpus usages compared to a random baseline. We predict a large number of unassigned usages that can be used to update WordNet's and SO's sense inventory in the near future. The models show different behavior on modern and historical data. One possible reason for this could be noise in the historical data, which worsens the model performance.

Based on the results of our experiments, we believe it is possible that our approach can be used in the near future to provide practical support in dictionary maintenance. WiC models seem to provide a promising basis for solving the task in a few-shot scenario.

Our results have a some limitations, though. Due to the experimental setup, an unrestricted comparison between the two English annotation phases is not possible. Future experiments should conduct pre- and post-prediction annotation conditions as similar as possible in order to avoid this limitation. In English, manual analysis of the data shows that the architecture still has some weaknesses: Faulty multi-word detection is a significant problem that influences model predictions. Another limitation is that the authors were unable to analyze the Swedish data manually because they do not speak the language.

Acknowledgements

We would like to thank the University of Gothenburg, especially Emma Sköldberg, for making it possible for us to work with Svensk Ordbok. This work has in part been funded by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021)

References

Adesam, Y., Dannélls, D., and Tahmasebi, N. (2019). Exploring the quality of the digital historical newspaper archive kubhist. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, Copenhagen, Denmark, March 5-8, 2019.*, pages 9–17.

Alatrash, R., Schlechtweg, D., Kuhn, J., and Schulte im Walde, S. (2020). CCOHA: Clean Corpus of Historical American English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France. European Language Resources Association.

Allén, S. (1981). The lemma-lexeme model of the swedish lexical data base. In Bo, R., editor, *Modersmålet i fäderneslandet. Ett urval uppsatser under fyrtio år av Sture Allén.*, pages 268–278. Meijerbergs arkiv för svensk ordforskning 25.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., and Pinkal, M. (2009). Framenet for the semantic analysis of german: Annotation, representation and automation. *Multilingual FrameNets in Computational Lexicography: methods and applications*, 200:209–244.

Cassotti, P., Siciliani, L., DeGemmis, M., Semeraro, G., and Basile, P. (2023). XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Chen, H., Xia, M., and Chen, D. (2021). Non-parametric few-shot learning for word sense disambiguation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1774–1781, Online. Association for Computational Linguistics.

Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7:121–157.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Diday, E. and Simon, J. C. (1976). *Clustering Analysis*, pages 47–94. Springer Berlin Heidelberg, Berlin, Heidelberg.

Erk, K. (2006). Unknown word sense detection as outlier detection. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 128–135, New York City, USA. Association for Computational Linguistics.

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

Fellbaum, C. (2005). Wordnet and wordnets. encyclopedia of language and linguistics.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Harris, Z. S. (1954). Distributional structure. *Word*, 10:146–162.

Hu, R., Li, S., and Liang, S. (2019). Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology.* SAGE Publications.

Kumar, S., Jat, S., Saxena, K., and Talukdar, P. (2019). Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681, Florence, Italy. Association for Computational Linguistics.

Kågebäck, M. and Salomonsson, H. (2016). Word sense disambiguation using a bidirectional lstm.

Laicher, S., Kurtyigit, S., Schlechtweg, D., Kuhn, J., and Schulte im Walde, S. (2021). 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*, pages 192–202, Online. Association for Computational Linguistics.

Lau, J. H., Cook, P., McCarthy, D., Gella, S., and Baldwin, T. (2014). Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 259–270, Baltimore, Maryland. Association for Computational Linguistics.

Lau, J. H., Cook, P., McCarthy, D., Newman, D., and Baldwin, T. (2012). Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, page 24–26, New York, NY, USA. Association for Computing Machinery.

Martinc, M., Montariol, S., Zosa, E., and Pivovarova, L. (2020). Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the Web*

Conference 2020, WWW '20, pages 343—-349, New York, NY, USA. Association for Computing Machinery.

Miller, G. A. (1994). WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Nasiruddin, M., Schwab, D., Tchechmedjiev, A., Sérasset, G., and Blanchon, H. (2014). Word sense induction for lexical resource enrichment (induction de sens pour enrichir des ressources lexicales) [in French]. In *Proceedings of TALN 2014 (Volume 2: Short Papers)*, pages 598–603, Marseille, France. Association pour le Traitement Automatique des Langues.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

Navigli, R. and Crisafulli, G. (2010). Inducing word senses to improve web search result clustering. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 116–126, Cambridge, MA. Association for Computational Linguistics.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018a). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237, New Orleans, LA, USA.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018b). Deep contextualized word representations.

Pilehvar, M. T. and Camacho-Collados, J. (2019). 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 Pa-*

pers), pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Rachinskiy, M. and Arefyev, N. (2022). 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*, Dublin, Ireland. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks.

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval.* McGraw - Hill Book Company, New York.

Scarlini, B., Pasini, T., and Navigli, R. (2020). Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8758–8765.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., and Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

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

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

Språkbanken (downloaded in 2019). *The Kubhist Corpus, v2.* Department of Swedish, University of Gothenburg.

Tabasi, M., Rezaee, K., and Pilehvar, M. T. (2022). Exploiting language model prompts using similarity measures: A case study on the word-in-context task. In

Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 325–332, Dublin, Ireland. Association for Computational Linguistics.

Towell, G. and Voorhees, E. M. (1998). Disambiguating highly ambiguous words. *Computational Linguistics*, 24(1):125–145.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

Véronis, J. (2004). Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language*, 18(3):223–252. Word Sense Disambiguation.

Weaver, W. (1949/1955). Translation. In Locke, W. N. and Boothe, A. D., editors, *Machine Translation of Languages*, pages 15–23. MIT Press, Cambridge, MA. Reprinted from a memorandum written by Weaver in 1949.

Yarowsky, D. (1994). Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 88–95, Las Cruces, New Mexico, USA. Association for Computational Linguistics.

Appendix

$F_{0.5}$-score								
	English				**Swedish**			
	DEFAULT		**SUBST.**		**DEFAULT**		**SUBST.**	
	COS	**SPR**	**COS**	**SPR**	**COS**	**SPR**	**COS**	**SPR**
E0	0.480	0.511	0.456	0.508	0.416	0.414	0.450	0.434
E1	0.540	0.564	0.548	0.577	0.431	0.439	0.442	0.444
E2	0.566	0.543	0.550	0.580	0.422	0.426	0.451	0.445
E3	0.518	0.525	0.515	0.526	0.421	0.418	0.431	0.431
E4	0.572	0.555	0.554	**0.585**				
G0	0.298	0.295	0.305	0.297	0.305	0.296	0.307	0.290
G1	0.276	0.256	0.269	0.251	0.507	0.514	0.479	0.482
G2	0.230	0.239	0.245	0.239	0.478	0.491	0.441	0.455
G3	0.252	0.245	0.270	0.249	0.534	**0.537**	0.506	0.524

Table 13: Results of cross-validation by maximizing $F_{0.5}$ score

$F_{0.1}$-Score								
	English				**Swedish**			
	DEFAULT		**SUBST.**		**DEFAULT**		**SUBST.**	
	COS	**SPR**	**COS**	**SPR**	**COS**	**SPR**	**COS**	**SPR**
E0	0.373	0.332	0.459	0.407	0.414	0.356	0.275	0.283
E1	0.765	0.646	0.826	0.806	0.414	0.356	0.275	0.283
E2	0.772	0.638	0.828	0.741	0.422	0.431	0.275	0.283
E3	0.902	0.885	0.664	0.636	0.414	0.356	0.275	0.283
E4	0.795	0.682	**0.850**	0.786				
G0	0.198	0.239	0.218	0.235	0.133	0.162	0.106	0.145
G1	0.127	0.121	0.133	0.128	0.772	**0.804**	0.521	0.746
G2	0.184	0.179	0.170	0.180	0.700	0.376	0.637	0.526
G3	0.193	0.20	0.182	0.177	0.692	0.491	0.609	0.548

Table 14: Results of cross-validation by maximizing $F_{0.1}$-Score