# Automatic Non-recorded Sense Detection for Swedish through Word Sense Induction with fine-tuned Word-in-Context models

**Dominik Schlechtweg[α], Emma Sköldberg[β], Shafqat Mumtaz Virk[β], James White[α], Simon Hengchen[γ]**

$\alpha$ University of Stuttgart, first.last@ims.uni-stuttgart.de
$\beta$ University of Gothenburg, first.last@svenska.gu.se
$\gamma$ Université de Genève & iguanodon.ai, first.last@unige.ch

**Abstract**

Finding non-recorded senses is important for dictionary maintenance, where using automatic methods helps reduce manual efforts. We use automatic Word Sense Induction (WSI) to compare recorded sense numbers among a sample of headwords in a comprehensive Swedish monolingual dictionary with induced sense numbers for the same words in a Swedish corpus. We propose this as a simple technique to find words to prioritize for post-hoc manual checks, which can be done in a simple Online-User-Interface bypassing the need for programming knowledge. We perform a thorough manual evaluation of the proposed methodology enabling us to show statistically that using automatic WSI increases the odds of finding non-recorded senses compared to a random selection of words. We further (i) evaluate predictions according to potential inclusion in the dictionary providing strong evidence for usefulness in practical lexicography, and (ii) analyze model predictions in-depth to point towards future improvements. We, finally, integrate lessons learned from our analysis into a large-scale prediction effort, providing the first high-quality large-scale WSI predictions for Swedish. These are a valuable resource for future research in Swedish lexicography.

**Keywords:** Non-recorded Sense Detection; Word Sense Induction; Word-in-Context; DURel; Swedish

## 1. Introduction

Identifying non-recorded senses is crucial for dictionary maintenance, i.e., keeping the dictionary up-to-date, and using automatic methods helps minimize manual efforts among lexicographers (Lau et al., 2012). We use automatic Word Sense Induction (WSI) to compare the recorded sense number among a sample of headwords in the *Svensk ordbok utgiven av Svenska Akademien* (i.e, the Contemporary Dictionary of the Swedish Academy, henceforth SO, 2021) to the induced sense number of the same words in corpus samples from the Sweden Television corpus in Korp/Språkbanken Text (SVT). This simple technique helps prioritize headwords for manual checks and can be carried out in the DURel tool without programming knowledge (Schlechtweg et al., 2024a), an Online-User-Interface allowing to upload corpus usages and to cluster them automatically. We build on previous

work (Sköldberg et al., 2024; Sander et al., 2024), introducing this idea, but extending it in various ways:

75. We add a manual evaluation enabling us to show statistically that using automatic WSI increases the odds of finding non-recorded senses against a random selection of headwords. This result is reproduced across two rounds of experiments.
76. We evaluate predictions according to potential inclusion in the SO dictionary providing strong evidence for usefulness in practical lexicography.
77. We analyze model predictions in-depth to point to future model improvements.
78. We integrate lessons learned from our analysis into a large-scale detection effort for Swedish headwords, providing the first high-quality, large-scale sense cluster predictions for Swedish. These predictions constitute a significant contribution to future research in Swedish lexicography.

Our main result is that sense clusters induced by automatic WSI methods based on state-of-the-art Word-in-Context models (Cassotti et al., 2023) have significant potential to be used as signals in dictionary maintenance. We provide a systematic, simple and general methodology to find non-updated dictionary entries that can be adopted by dictionary-makers without programming knowledge. Through iterative improvement of preprocessing corpus samples (filtering out e.g. proper nouns and short contexts) we considerably improve detection quality. Sense granularity is an adjustable parameter in our methodology, which we find to align well with SO's sense distinctions after tuning the parameter on a set of expert lexicographer annotations. We found various non-covered senses which will be included into the dictionary at some point. Furthermore, some new usages will be monitored to see if they become more established in Swedish in the future. In other cases, our methodology points out definitions that need to be supplemented, examples that need to be added, or idioms that should be recorded.

While some parts of our approach are novel in the field of lexicography (see Sec. 2.2), our aim in this work is not to identify optimal WSI models for the task of Non-recorded Sense Detection (NSD), but to provide a more in-depth evaluation of a particular approach in a realistic, large-scale setting.

## 2. Related work

### 2.1 Dictionary maintenance

Applied lexicographic work involves creating new dictionaries and other types of lexical resources. However, it also involves the maintenance and development of already existing resources. Obtaining financial support for the revision of dictionaries etc. can be challenging, even though it is resource-intensive. The work does not just involve updating the list of headwords. Information categories in the entries can be added, and the existing content, like language examples, may need to be updated to keep the dictionary up to date. Making a previously printed dictionary increasingly digital, for instance, by adding new functions, is another example of lexicographic maintenance work. Using language technology methods and tools to support maintenance work is highly advantageous for numerous reasons, not least because of the cost savings in the dictionary projects (see e.g. Cook et al. 2013; Grundy & Rawlinson 2015:561–578; Nilsson 2025:507ff.).

### 2.2 Non-recorded sense detection

Erk (2006) gives the first systematic approach to automatic NSD treating the problem as a binary classification task: Word usages that are not covered by any entry in a sense inventory need to be assigned label 1 while covered usages need to be assigned label 0. Erk proposes two types of models, one relying on a Word Sense Disambiguation (WSD) classifier exploiting word sense definitions from the dictionary, and one relying purely on word usages and an outlier detection model, which is more related to WSI because sense definitions are ignored. Lautenschlager et al. (2024) use a more WSD-like approach with English and Swedish data being the first to use modern embedding models for the task. Recently, a shared task on NSD was organized mixing aspects of WSD and WSI (Fedorova et al. 2024).

Several earlier studies have explored the detection of non-recorded senses from a lexicographic perspective. Notable examples include Cook et al. (2013) and Cook et al. (2014) which focus on English, Nimb et al. (2020), which examines Danish, and Cheilytko & von Waldenfels (2024a, 2024b), which investigates Ukrainian. Although some of these studies use WSI models, they mostly compare word usage across corpora, i.e., a dictionary is not directly compared to. Further, none of them apply state-of-the-art Word-in-Context models, and they do not apply their methods in large-scale settings.

Related tasks are Novel Sense Detection (Lau et al. 2012; Jana et al. 2020) and Lexical Semantic Change Detection (Schlechtweg et al. 2020; Kurtyigit et al. 2021), with the difference from NSD being that they do not assume the existence of a dictionary, but rely on the comparison of corpora.

#### 2.2.1 Our previous experiments

In our previous experiments (Sköldberg et al. 2024, Sander et al. 2024), we approached the task of NSD in a more WSI-like fashion: We sampled word usages for dictionary headwords from recent corpora in various languages and clustered these into senses using a computational model (cf. Schlechtweg et al. 2024b). We then compared induced sense numbers to sense numbers in the reference dictionaries and treated any deviations as signals for non-recorded senses. We then manually analyzed our predictions to evaluate the effectiveness of the approach. Up to roughly 50% of the detected words were found to exhibit non-recorded senses. Hence, we concluded that our approach has potential to be used in dictionary maintenance and conducted further experiments, as described in this study.

## 3. Data

### 3.1 The Swedish dictionary SO

In our work, we rely on the semantic descriptions in the comprehensive Swedish Academy's defining dictionary, SO, which is developed at Språkbanken Text at the University of Gothenburg. SO is available as app and on the dictionary portal at svenska.se. The latest edition was published in 2021, and the next update of the dictionary is planned for early 2026. The dictionary, which covers contemporary Swedish, contains approximately 65,000 headwords. The semantic information is crucial in SO and it includes approximately

100,000 different senses, in form of main senses and subsenses, in total. The difference between the different senses is subtle. Traditionally, the SO lexicographers' work on finding new senses has been based on manual efforts. However, language technology-based methods can both streamline and (further) improve the quality of this part of the lexicographical work (see more about the meanings descriptions in the dictionary in, e.g., Sköldberg et al. 2024; Sköldberg et al. 2025).

In order to transform the production database of SO2021 into an electronic format that is easier to work with, we extracted all headwords and relevant information (parts of speech, senses, glosses, sub-glosses, years, and example sentences) into a nested JSON. This format allows for easy and fast processing of the dictionary data.[1] We extracted a total of 39,146 single-sense headwords from the SO dictionary dump. Among these, 29,700 are nouns, 5,427 are adjectives, and 4,019 are verbs. Other parts-of-speech were excluded. For our main study described in Sec. 6.2, we randomly selected a subset of 1175 headwords. To ensure a balanced representation of parts of speech, the selection was made proportionally based on the distribution of nouns, adjectives, and verbs in the full set.

### 3.2 The SVT corpus

The SVT corpus available through Språkbanken Text is a collection of Swedish written news and news articles published by Sveriges Television (SVT), Sweden's national public broadcaster. The corpus consists of 21 sub-corpora, encompassing a total of 240,393,329 tokens and 15,991,049 sentences covering the period from 2004 to 2021. It is linguistically annotated with lemmatization, part-of-speech tags and syntactic dependency relations. The annotations are produced using Sparv – Språkbankens annotation pipeline – which relies on a number of external as well as in-house built annotation tools (see Hammarstedt et al. 2022). The corpus offers a rich and valuable resource for research in natural language processing and empirical linguistic studies.

## 4. Task

Lautenschlager et al. (2024) define the task of NSD as a binary classification task on the usage level: Let $U = \{u_1, u_2 u_n\}$ be a set of word usages of a word $w$. Let $\phi = \{\phi_1, \phi_2 \phi_k\}$ be a set of predefined senses and f : U $\rightarrow$ $\phi$ be a mapping from usages to senses given by a lexical resource. Let further g : U $\rightarrow$ {0, 1} be a mapping such that g(u) = 1 iff f is undefined for u and g(u) = 0 otherwise. The task is to find mapping g. That is, word usages that are not covered by any entry in the sense inventory need to be assigned label 1 while covered usages need to be assigned label 0. We solve a related, but similar task on the word level: Let h : w $\rightarrow$ {0, 1} be a mapping such that h(w) = 1 iff g(u) = 1 for any u $\in$ U and 0 otherwise. The task is to find mapping h. That is, we want to label a word with 1 if it has *any* usage with an unknown sense.

[1] The code, along with a description of the data structure, is freely available under a permissive licence: https://github.com/iguandon-ai/SO-extract-db . See the "`data-structure.md`" file for an example on the data structure.

## 5. Models and Tool

We model mapping h from Sec. 4 by assigning words with more than one predicted sense (>1-cluster) to class 1 and 0 otherwise (1-cluster). As all our target words are guaranteed to have only one known sense in the dictionary (Sec. 3.1), any word with more than one sense in the corpus sample can be assumed to have at least one unknown sense.

Sense clusters are inferred with the DURel annotation tool (Schlechtweg et al., 2024a).[2] It provides functionalities to upload sets of word usages for a particular target word and to annotate the semantic proximity between these usages either with humans or computers. This information (usages with their proximities) can then be visualized in a graph and clustered into sets of high-proximity usages, which can be interpreted as word senses.

The tool allows to inspect the graph and to adjust multiple annotation, cluster and visualization parameters. In this way, we can easily infer lexicographically relevant semantic information such as the inferred number of senses or their changes over time. Find an example of a DURel graph concerning the usage of the noun *skyltfönster* ('shop window') in the SVT corpus in Fig. 1.

DURel offers multiple computational annotators for semantic proximity based on state-of-the-art pre-trained Word-in-Context models (Pilehvar & Camacho-Collados, 2019; Yadav & Schlechtweg, 2025). We rely on the model XL-LEXEME which was optimized for Lexical Semantic Change Detection (Cassotti et al., 2023). This model maps each word usage onto a contextualized distributional-semantic vector representation and then uses cosine similarity to estimate the semantic proximity between two usage vectors.

Fig. 1 visualizes usages of the word *skyltfönster* 'shop window' as nodes while edge weights are given by the cosine similarity between usages annotated with the XL-LEXEME model. The graph was clustered using Correlation Clustering (Schlechtweg et al., 2020). In SO, the noun *skyltfönster* has one sense: "(avgränsad yta innanför) stort butiksfönster för skyltning", i.e. '(enclosed area inside) large shop window for display'. However, DURel highlights two semantic clusters in the corpus, represented by blue and orange dots and users can click on individual dots in the DURel front-end interface to view the corresponding corpus samples. In this case, among the blue dots there is the sample "En bil körde av misstag in i skyltfönstret på en restaurang på S:t Persgatan i Uppsala på torsdagsförmiddagen." 'A car accidentally drove into the shop window of a restaurant on S:t Persgatan in Uppsala on Thursday.' Among the orange dots, there is the sample "Arenan och elithockeyn är ett bra skyltfönster för hela Oskarshamn." 'The arena and elite hockey are a great showcase for the entire Oskarshamn.' In this case, DURel has identified two distinct senses of the word in the corpus, with the latter being metaphorical. The figurative use will most likely also be described in an updated SO.

## 6. Experiments

### 6.1 Pilot study

In our previous study (Sec. 2.2.1), we performed two experiments: In the first one, we tuned clustering parameters on sense annotations of 18 manually selected SO-headwords.

[2] https://durel.ims.uni-stuttgart.de/

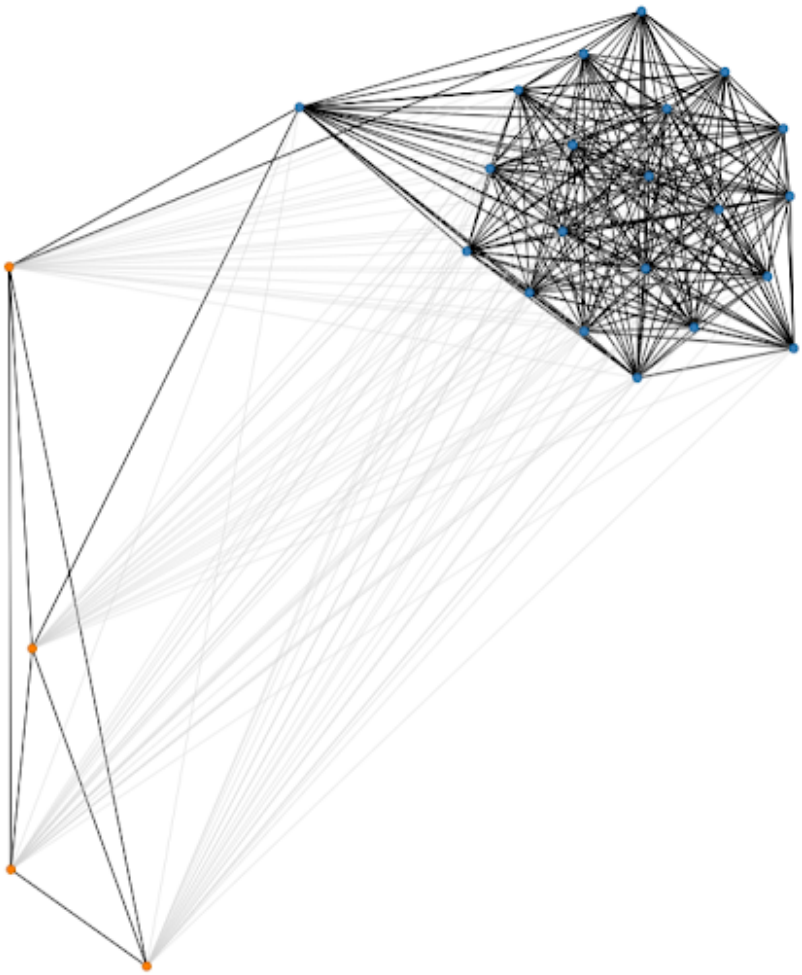

Figure 1: The noun *skyltfönster* ‘shop window’, based on 25 corpus samples from the SVT corpus, forming two semantic clusters in DURel.

Our second experiment can be seen as a smaller pilot study preparing the one we present in this article. We randomly selected 281 entries in SO with only one sense (nouns, adjectives, verbs). All these entries have at least 25 occurrences in the SVT corpus. Uses were extracted from the corpus and DURel clusters were created. The pilot study resulted in some improvements in SO. The annotation tool predicted that 66 headwords, i.e., about a quarter of the selected ones, have 2–6 senses. Some examples are: *armkrok* ‘arm in arm’, *brottningsmatch* ‘wrestling match’, *genrep* ‘dress rehearsal’, *guldmakare* ‘alchemist’, *lydnad* ‘obedience’, *handplocka* ‘pick by hand’, *minera* ‘mine’, *party* ‘party’ and *slutspurt* ‘final sprint’. New senses for these words have been added to the SO database and they will be made public in the next SO update (see Sköldberg et al. 2024 on this study).

However, we also gained more general insights from the pilot study. SO-lexicographers manually assessed clusters and language samples. A comparison between the lexicographers’ and DURel’s analyses showed that the results of the pilot study are solid (see also Sec. 6.3). However, the method and the examined data can be further improved: We found that we can avoid many difficult-to-interpret cases by requiring more context in the corpus samples. Furthermore, we avoid many problematic names if we exclude corpus samples where the word in focus starts with a capital letter.

### 6.2 Main study

We then randomly selected 1175 entries (nouns, adjectives, verbs) with one sense in SO (i.e. almost 4 times as many entries as in the pilot study). 25 uses of each word were extracted from the SVT corpus.[3] Based on insights from the pilot study, certain filtering criteria were applied: Usages were considered duplicates and removed if the five tokens on both sides of the headword were identical across usages. Additionally, any usage with fewer than 10 tokens[4] was considered to provide insufficient context and was excluded from the dataset. Finally, DURel clusters were generated exactly as in the pilot study.

For 956 of the 1175 entries, DURel formed 1 cluster. However, a total of 219 of the selected headwords have, according to DURel, 2–6 senses. This corresponds to approximately one fifth of the cases. We now discuss a number of examples. A rigorous manual error analysis of the model predictions is given in Sec. 6.3.

**Headwords with 1 cluster by DURel**: The majority of the examined headwords (in total 956) have only 1 cluster in DURel. Among them you find *antibakteriell* 'antibacterial', *distansminut* 'nautical mile', *sakristia* 'sacristy', *smaksätta* 'flavour', *årsinkomst* 'annual income' and *beachvolleyboll* 'beach volleyball' (see Fig. 2).

**Headwords with 2 or more DURel clusters**: As already mentioned, in approximately one fifth of the cases, DURel forms 2 or more clusters. Among these cases you find the noun *botemedel*. According to SO, it is understood as 'a means to eliminate a problem, often a disease, but also more generally.' For this word, DURel has created 2 clusters (see Fig. 3).

Among the blue dots, there are corpus examples like "Narkolepsi är en kronisk sjukdom och det finns alltså inget botemedel." ('Narcolepsy is a chronic illness, and therefore, there is no cure.') However, among the orange dots, there are more figurative examples, like "Men [politikern] Anna Starbrink tror inte på att politisk styrning är botemedlet mot vårdkrisen." ('But [the politician] Anna Starbrink does not believe that political governance is the remedy for the healthcare crisis.')

Here are additional examples of entries where SO has one sense, while DURel indicates that there are three or more senses:

- 3 clusters: *frysbox* 'freezer', *lätthet* 'ease', *obscen* 'obscene', *tråla* 'trawl'
- 4 clusters: *anrop* 'call', *gena* 'make a shortcut', *pika* 'to peak', *resumé* 'summary'
- 6 clusters: *inaktiv* 'inactive'

In e.g. the case of the verb *tråla* 'trawl', the literal sense 'to fish with a trawl' is captured in one major cluster (with 21 samples). The second cluster includes one sample and it has to do with policemen trawling after a wanted car (cf. the sense of 'to fine-tooth comb').

[3] Words with fewer corpus usages were ignored. The choice of 25 usages was made after some experimentation. Since SVT is a relatively small corpus and certain headwords did not have many occurrences, we tested different options and determined that 25 usages per headword was the most suitable compromise between remaining headwords and processing load for the computational models.

[4] The extracted usages were restricted to single sentences as the corpus is segmented at the sentence level.

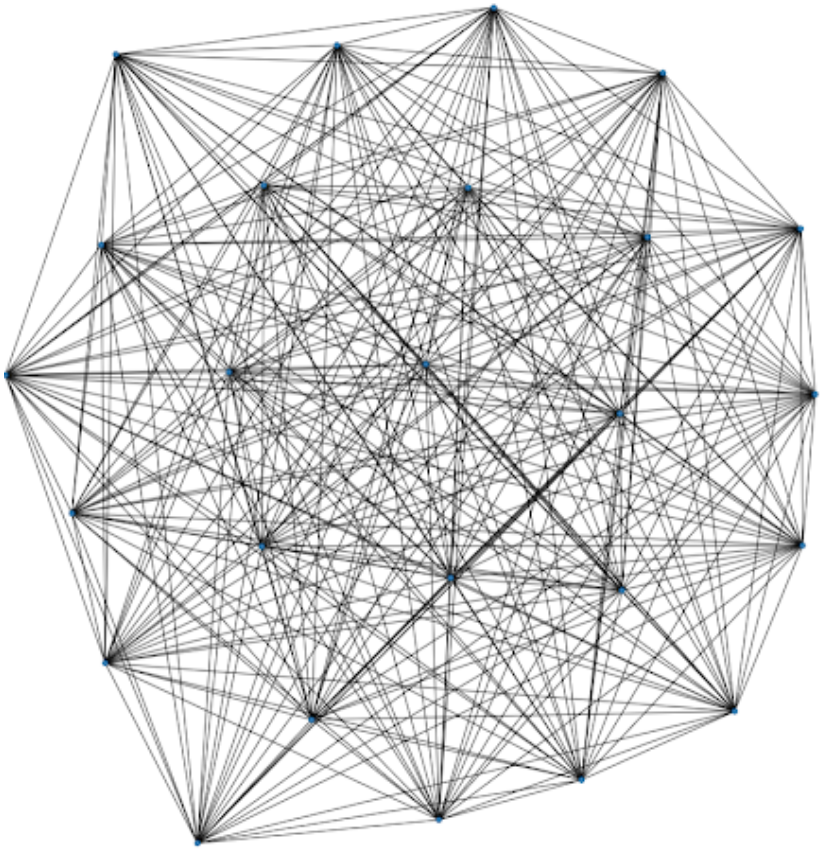

Figure 2: The noun *beachvolleyboll* 'beach volleyball', based on 25 corpus samples from the SVT corpus, forming one semantic cluster in DURel.

The third cluster involves three samples of trawling on the internet. The two figurative senses of the word captured in the second and the third cluster are (still) not recorded in SO.

Another example is the adjective *obscen* 'obscene', with the following meaning description in SO: "präglad av ohöljd sexualitet" ('marked by overt sexuality'). DURel has formed three clusters based on the SVT samples. One of them (with 21 of 25 samples) includes many cases of the word combination *obscena gester* 'obscene gestures'. The second one, based on three samples, includes e.g. the word combination *obscena bonusar* 'obscene bonuses'. Also in these cases, the word involves actions or similar that can be perceived as provocative or distasteful but the semantic feature 'sexuality' is not present in the same way. The third and last cluster involve a sample where the sense of the target word is unclear.

### 6.3 Manual analysis and lexicographic annotation

SO lexicographers conducted manual analysis of the two groups of interest (1-cluster vs. >1-clusters). For this, we randomly sampled 28 words from the 1-cluster group and all 66 words from the >1-cluster group in the pilot study for analysis. We further sampled 21 and 153 words from the respective groups in the main study. These were then manually analyzed involving an analysis (and comparison) of the existing meaning descriptions of the headwords in SO, DURel's clusters, the content of the word usages in the individual

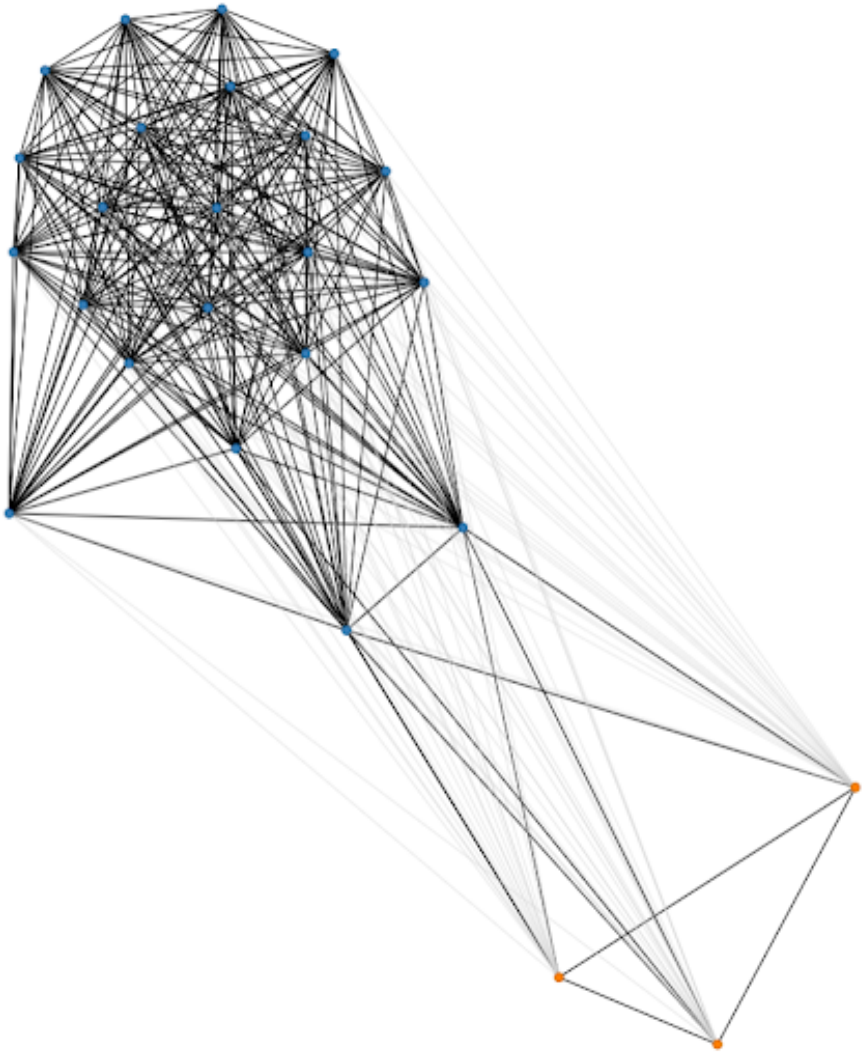

Figure 3: The noun *botemedel* 'remedy', based on 25 corpus samples from the SVT corpus, forming two semantic clusters in DURel.

corpus samples, and how the samples were distributed across the clusters. The main aim of the analysis was to determine whether a semantic variation in meaning was present among the samples that could be explained through the **presence of a non-recorded sense**. Note that such a presence does not necessarily imply an update in the dictionary, which is discussed in Sec. 6.4.

**Non-recorded senses**: The outcome of the manual analysis is shown in Tab. 1 and 2. For the pilot, the 1-cluster group has 1 non-recorded sense case out of 28 (4%) while the >1-cluster group has 30 cases with non-recorded senses out of 66 (45%). For the main study, the 1-cluster group has 0 non-recorded sense cases out of 21 (0%) while the >1-cluster group has 87 such cases out of 153 (57%). Hence, we can see that the >1-cluster group has a much larger proportion of non-recorded sense cases in both studies.

**Are cluster groups significantly different?** We now perform a statistical analysis to test whether the observed difference in non-recorded senses is significant, i.e., is likely not a result of chance and can thus be attributed to meaningful semantic information induced by the computational model. For this, we employ Fisher's exact test for independence between

Table 1: 2×2 contingency table for cluster number vs. non-recorded sense presence in the *pilot study*.

| **Item** | **1-cluster** | **>1-cluster** | **Row total** |
|---|---|---|---|
| non-recorded | 1 | 30 | 31 |
| recorded | 27 | 36 | 63 |
| **Column total** | **28** | **66** | **94** |

Table 2: 2×2 contingency table for cluster number vs. non-recorded sense presence in the *main study*.

| **Item** | **1-cluster** | **>1-cluster** | **Row total** |
|---|---|---|---|
| non-recorded | 0 | 87 | 87 |
| recorded | 21 | 66 | 87 |
| **Column total** | **21** | **153** | **174** |

two categorical variables (Fisher, 1922). In our case these variables are given by the cluster number (1-cluster vs. >1-cluster) vs. non-recorded sense presence (present vs. not present). For the pilot study, we represent these two variables in a 2×2 contingency table in Tab. 1. For this purpose, Fisher's exact test (one-tailed) tests the probability of the null hypothesis that words with one cluster are equally or more likely to have non-recorded senses than words with more clusters. For Tab. 1, the test gives a probability of 0.00002412. Hence, we can reject the null hypothesis assuming any standard significance thresholds such as <0.01, and conclude that the alternative hypothesis holds, i.e., words with 1 cluster are less likely to have non-recorded senses. The low probability reflects the strong observed effect size: 4% vs. 45%. Additionally, the 1-cluster group is much more frequent than the >1-cluster group in the full data that were clustered in the pilot study (215 versus 66 words respectively; 77% of all words). Hence, by inspecting only the >1-cluster group instead of the full data we significantly increase the chance to find non-recorded senses compared to a random selection. For Tab. 2, the test gives a probability of 0.00000012. Hence, we can equally conclude that the alternative hypothesis holds for the main experiment. The effect size is even higher for the main study (0% vs. 57%). Also, the 1-cluster group is much more frequent than the >1-cluster group in the full data that were clustered (956 versus 219 words respectively; 81% of all words).

The proportions indicate that roughly every second word predicted by our approach actually has a non-recorded sense, and that almost all words excluded by our approach have no such sense. The increase in effect size between pilot and main study further indicates that our modifications of the model pipeline (see Sec. 6.1) did have a positive effect. Combined with the large proportion of the 1-cluster group observed in both studies (roughly 80%), we conclude that our approach will significantly increase the odds to find non-recorded senses over a random selection procedure.

**Prediction analysis**: We now report our observations from the manual analysis on each of the four groups from Tables 1 and 2: (i) Cases in >1-cluster indeed found to have a non-recorded sense and (ii) cases with 1-cluster indeed found not to have a non-recorded sense can be seen as successful model predictions confirming our assumption that cluster number indicates non-recorded senses. In contrast, (iii) cases in >1-cluster found not to have a non-recorded sense and (iv) cases with 1-cluster found to have one can be seen as model prediction errors as they contradict our assumption. We provide a representative example from each group below.

In group (i), we find the adjective *inaktiv* 'inactive', with the meaning description 'not active (especially referring to a person)' in the dictionary (see Sec. 6.2). From the language examples included in the entry, it is clear that the description refers to participation in community life, etc. By forming 6 clusters DURel indicates one or more new senses among the SVT samples besides the one recorded in SO, and this is confirmed by the lexicographers. For instance, SO (2021) lacks a subsense concerning physical inactivity and lack of exercise.

A target word that serves as a representative for group (ii) is the already mentioned *beachvolleyboll* 'beach volleyball' (see Sec. 6.2). DURel creates 1 cluster and there is just one sense, i.e., 'volleyball played on a sand surface (and often outdoors)' represented among the samples.

Group (iii) consists of cases where DURel creates two or more clusters, but there is only one sense among the corpus samples. For instance, for the noun *barnbarn* 'grandchild' with only one sense in SO, DURel creates two clusters. The reason for this may be that the target word is used in slightly different constructions in the SVT samples. In the larger cluster, the noun is almost exclusively found in the construction "someone's grandchild", while in the other cluster (with only one sample), the noun appears in the construction "the grandchild of someone". Information about both constructions is already included in SO, and the difference between the two uses does not, in this case, lead to different senses in the dictionary. More examples of this type are provided in Sköldberg et al. (2024:178).

And, finally, in group (iv), DURel gives only 1 cluster, but an analysis of the SVT samples shows that the target word has at least two senses. We have no such case in the main study. But, as mentioned in Sköldberg et al. (2024:178-179), e.g., the noun *minfält* 'minefield' from the pilot study has two senses in the corpus while DURel induces only one cluster. Among the examples, one finds both instances such as "minfälten längs frontlinjen" 'the minefields along the front line,' as well as "ett formidabelt minfält av ideologiska teorier" 'a formidable minefield of ideological theories'. Thus, in comparison to the semantic description in SO, there is also a figurative, non-recorded, sense in the corpus.

Group (i) and (iii) contain quite a few cases where one or more clusters involve idioms. For example, the entry *hjul* 'wheel' can be mentioned. DURel has created two clusters, one with 16 samples and one with 9 samples. The former contains literal uses of the word. The latter, however, mainly consists of samples with uses of the idiom *sätta käppar i hjulet/hjulen* (cf. put/throw a spanner in the works). This Swedish idiom is already described in SO. However, DURel has previously identified different variants of multiword expressions that are not already recorded in the dictionary (see Schlechtweg et al. 2024a:14 about *ha något i bagaget* 'to have something in the luggage').

In general, the headwords predicted to have 1 cluster in DURel (1-cluster group) have one sense in SO and the same meaning in the corpus examples. This indicates accurate semantics in these SO entries. The headwords predicted to have more than 1 cluster in DURel (>1-cluster group), are more heterogeneous: In roughly half of the predicted cases, we find a non-recorded sense. However, not all language samples are correctly placed in their respective clusters and the number of clusters and the number of senses do not always match. Moreover, there are still usages among the samples where the target word's meaning is ambiguous although these instances are less frequent in the main study compared to the pilot study. In addition, we mainly find older senses that should already have been included in SO from 2021 (but also few newer senses).

### 6.4 How likely are updates of the dictionary?

A decision on whether to add a new sense to SO is based, among other things, on how frequent and widespread the particular sense is in modern Swedish texts, specifically those that contain general language rather than specialized terminology. Sometimes, it is sufficient for the experienced lexicographers to look only at evidence in the SVT corpus to determine that a sense, beyond the one that is already recorded in the dictionary, is older and sufficiently established. Other times, e.g., when it comes to senses that result from recent borrowings from other languages, searches in additional corpora may be required to make this determination. Whether a main sense or a subsense is added follows the traditions applicable to the semantic description in the dictionary, which, as mentioned, is relatively detailed (see Sköldberg et al. 2024).

We inspected a first sample of words from groups (i) and (iv). From these, the dictionary entries of the above-mentioned *skyltfönster* 'shop window', *botemedel* 'remedy', *tråla* 'trawl', and *obscen* 'obscene' will likely be revised. Some other SO entries that most probably will receive (at least) new figurative subsenses are, e.g., *coacha* 'to coach', *huvudrätt* 'main course', *pantsätta* 'to pawn', *styvbarn* 'stepchild' and *thriller* 'thriller'. In addition, for an adjective like *ömsint* 'tender', an extended sense should be added in the dictionary. In SO, the adjective *ömsint* should be described in the same way as another adjective, *kärleksfull* 'loving,' which is used similarly. The main sense of *ömsint* should pertain to a person's characteristics. In addition, a new subsense covers (the result of) an action (cf. 'a tender parent' and 'a tender description').

## 7. Conclusions and further work

In this study we have solidified some of our previous results suggesting that WSI can be used effectively in the detection of non-recorded senses by simply comparing induced sense numbers to dictionary sense numbers. We performed a manual analysis to evaluate the predictions and conducted a statistical test clearly showing that the predictions significantly increase the chance to find non-recorded senses. We then integrated the lessons learned from the manual analysis into a large-scale experiment showing that we are able to further increase the effect size of the prediction procedure. In total we created computational predictions for roughly 1,500 target words, which we make available for follow-up studies.[5] The manual analysis shows that the method presented in this article enables the SO lexicographers to concentrate on entries in the dictionary that are more

[5] https://doi.org/10.5281/zenodo.15850761

likely to lack one or more senses. Several well-established senses (both main senses and subsenses) not covered by SO as of May 15th, 2025, have been identified and included in the lexical database. The interdisciplinary work underlying the experiments also contributes to increased awareness and critical examination among both practicing lexicographers and language technologists of how lexicographic work is conducted and what it entails.

Our modeling approach uses WSI methods and thus ignores the abundance of available word sense definitions from the dictionary. These could possibly serve as valuable information in the prediction process as they allow for a direct comparison between dictionary entries and corpus samples. Hence, in the future, we would like to combine our method with WSD models incorporating these definitions (e.g. Kokosinskii et al. 2024). Further, we would like to improve the matching of corpus usages to multi-word expressions like idioms or phrasal verbs in SO and include the size of clusters (rather than only the cluster number) into the detection process of update candidates in the dictionary. Also, more corpus evidence may need to be studied, as well as content in other corpora. Moreover, we focused only on monosemous content headwords. We would like to repeat the experiments by including e.g. function words in the analysis, incorporating additional information from the dictionary beyond the definitions, and examining words with two or more senses. We would also like to analyze the cluster groups in more detail. For example, we could examine whether the headwords with more DURel clusters have anything in common.

## 8. Acknowledgments

This work has in part been funded by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021).

## 9. References

Cassotti, P., Siciliani, L., de Gemmis, M., Semeraro, G., & 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, (pp. 1577–1585). Association for Computational Linguistics.

Cheilytko, N. & von Waldenfels, R. (2024a). Semantic change and lexical variation in Ukrainian with vector representations and LLM. In: Krek, S. (Ed.), Book of Abstracts of the Workshop Large Language Models and Lexicography, 1–5.

Cheilytko, N. & von Waldenfels, R. (2024b). Word Embeddings for Detecting Lexical Semantic Change in Ukrainian. In: Despot, K. Š., Ostroški Anić, A. & Brač, I. (Eds.), Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress, 231–243.

Cook, P., Lau, J. H., Rundell, M., McCarthy, D., & Baldwin, T. (2013). A lexicographic appraisal of an automatic approach for detecting new word senses. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, & M. Tuulik (Eds.), Proceedings of eLex 2013 conference (pp. 49–65). Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

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

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

Fedorova, M., Mickus, T., Partanen, N., Siewert, J., Spaziani, E. & Kutuzov, A. (2024). AXOLOTL'24 Shared Task on Multilingual Explainable Semantic Change Modeling. In: Tahmasebi, N., Montariol, S., Kutuzov, A., Alfter, D., Periti, F., Cassotti, P. & Huebscher, N. (eds.), Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change (pp. 72–91). Bangkok, Thailand: Association for Computational Linguistics.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222(594–604), 309–368.

Grundy, V. & Rawlinson, D. (2015). The practicalities of dictionary production; planning and managing dictionary projects; training of lexicographers. In: Durkin, P. (ed.), The Oxford Handbook of Lexicography. Oxford: Oxford University Press, 561–578.

Hammarstedt, M., Schumacher, A., Borin, L., & Forsberg, M. (2022). Sparv 5 User Manual.

Jana, A., Mukherjee, A. & Goyal, P. (2020). Network measures: A new paradigm towards reliable novel word sense detection. Information Processing & Management, 57(6), 102173.

Kokosinskii, D., Kuklin, M. & Arefyev, N. (2024). Deep-change at AXOLOTL-24: Orchestrating WSD and WSI models for semantic change modeling. In: Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change, Bangkok, Thailand. Association for Computational Linguistics, 168–179.

Kurtyigit, S., Park, M., Schlechtweg, D., Kuhn, J. & Schulte im Walde, S. (2021). Lexical semantic change discovery. arXiv preprint arXiv:2106.03111.

Lau, J. H., Cook, P., McCarthy, D., Newman, D., & Baldwin, T. (2012). Word sense induction for novel sense detection. In: W. Daelemans (Ed.), Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 591–601). Avignon, France: Association for Computational Linguistics.

Lautenschlager, J., Sköldberg, E., Hengchen, S., Schlechtweg, D. (2024). Detection of Non-recorded Word Senses in English and Swedish. arXiv eprint. Available at https://arxiv.org/abs/2403.02285.

Nilsson, P. (2024). Report on the revision of the Swedish Academy Dictionary – and the search for "old neologisms". In: Despot, K. Š., Ostroški Anić, A. & Brač, I. (eds.), Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress, 8–12 October 2024, Cavtat, Croatia, 507–522.

Nimb, S., Sørensen, N. H. & Lorentzen, H. (2020). Updating the dictionary: Semantic change identification based on change in bigrams over time. Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave, 8(2), 112–138.

Pilehvar, M. T. & 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 Papers), Minneapolis, Minnesota. Association for Computational Linguistics, 1267–1273.

Sander, P., Hengchen, S., Zhao, W., Ma, X., Sköldberg, E., Virk, S. & Schlechtweg, D. (2024). The DURel Annotation Tool: Using fine-tuned LLMs to discover non-recorded senses in multiple languages. In: Proceedings of the Workshop on Large Language Models and Lexicography at 21st EURALEX International Congress Lexicography and Semantics.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H. & 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.

Schlechtweg, D., Virk, S., Sander, P., Sköldberg, E., Theuer Linke, L., Zhang, T., Tahmasebi, N., Kuhn, J. & Schulte Im Walde, S. (2024a). The DURel Annotation Tool: Human and Computational Measurement of Semantic Proximity, Sense Clusters and Semantic Change. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, (pp. 137–149). Association for Computational Linguistics.

Schlechtweg, D., Zamora-Reina, F. D., Bravo-Marquez, F. & Arefyev, N. (2024b). Sense through time: diachronic word sense annotations for word sense induction and lexical semantic change detection. Language Resources and Evaluation, 1–35.

Sköldberg, E., Virk, S., Sander, P., Hengchen, S. & Schlechtweg, D. (2024). Revealing semantic variation in Swedish using computational models of semantic proximity - Results from lexicographical experiments. In: Despot, K. Š., Ostroški Anić, A. & Brač, I. (Eds.), Proceedings of the 21st EURALEX International Congress Lexicography and Semantics, 169–182.

Sköldberg, E., Blensenius, K. & Holmer, L. (2025). SO: the Swedish contemporary dictionary. In: Dannélls, D., Blensenius, K. & Borin, L. (eds.), Sixty Years of Swedish Computational Lexicography. (Digital Linguistics, 3). De Gruyter, 53–82.

SO May 15th, 2025: Svensk ordbok utgiven av Svenska Akademien ['The Contemporary Dictionary of the Swedish Academy']. (2021). 2nd edition. Retrieved May 15, 2024, from https://svenska.se/.

SVT (Sweden Television) corpus. Retrieved May 15, 2024, from Språkbanken's word research platform Korp, available at https://spraakbanken.gu.se/korp

Yadav, S. & Schlechtweg, D. (2025). XL-DURel: Finetuning Sentence Transformers for Ordinal Word-in-Context Classification. ArXiv preprint. https://arxiv.org/abs/2507.14578.