University of Stuttgart
Germany

UNIVERSITÉ DE GENÈVE

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

November 18, 2025

Dominik Schlechtweg, Emma Sköldberg, Shafqat Virk, James White, Simon Hengchen

eLex 2025

IGUANODON AI

UNIVERSITY OF GOTHENBURG

Introduction

- **motivation**: dictionaries need to be maintained
- **task**: find incomplete dictionary entries
- **data**: the monolingual Swedish dictionary SO
- **solution**: compare dictionary sense number to corpus sense number
- **method**: automatic word sense induction with word-in-context models
- **contributions**:
  1. accurate, systematic, simple and general methodology
  2. manual and statistical analysis
  3. realistic and large-scale setting

Related work

- **non-recorded sense detection** (Cook et al., 2014; Erk, 2006; Fedorova et al., 2024)
  - binary classification problem
  - combines aspects of word sense induction and disambiguation
- **our previous studies** (Sander et al., 2024; Sköldberg et al., 2024)
  - word sense induction
  - small-scale Swedish/English/German data
  - ~50% had non-recorded senses

Task

- **usage-level**: Given a usage u and a set of sense descriptions S, decide whether the sense of u is covered by S.
- **lemma-level**: Given a set of usages U and a set of sense descriptions S, decide whether any of the usages $u \in U$ is not covered by S.

Data

- SO = the comprehensive **Swedish Academy**'s defining dictionary
  - developed at Språkbanken Text, University of Gothenburg
  - available as app and on the dictionary portal Svenska.se.
  - latest edition: 2021; next update: early 2026
  - covers **contemporary** Swedish (about 65,000 headwords in total; about 39,000 of them (60%) are reported to have only one sense)
- SVT corpus = Swedish written **news** and news articles published by Sweden's national public broadcaster
  - available through Språkbanken Text
  - 18 subcorpora from 2004-2021 (around 218 million tokens)

Model & Tool

1. upload target word usages to DURel tool[1] (Schlechtweg et al., 2024)
2. automatically annotate all usage pairs with **semantic proximity** score
3. represent scores as edges in weighted **graph**
4. automatically **cluster** edges into sense clusters
5. treat all words with more than one cluster as having usages of a **non-recorded** sense

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

Corpus examples for *botemedel* 'remedy'

(A) Medicinen stoppar RNA-virus från att föröka sig – och är snarare en bromsmedicin än ett **botemedel** .
*'The medicine stops RNA viruses from multiplying – and is more of a brake drug than a **cure**.'*

(B) En grupp forskare vid Imperial College i London kan ha upptäckt ett nytt **botemedel** mot åk- och sjösjuka .
*'A group of researchers at Imperial College London may have discovered a new **cure** for motion sickness.'*

(C) IFK Göteborg hade inget **botemedel** mot (...) Aliou Badji borta mot Djurgården .
*'IFK Göteborg had no **remedy** against (...) Aliou Badji during their away game against Djurgården.'*

. . . (25 usages in total)

Example pair for *botemedel* ‘remedy’

(A) Medicinen stoppar RNA-virus från att föröka sig – och är snarare en bromsmedicin än ett **botemedel** .
‘The medicine stops RNA viruses from multiplying – and is more of a brake drug than a ***cure****.’*

(B) En grupp forskare vid Imperial College i London kan ha upptäckt ett nytt **botemedel** mot åk- och sjösjuka .
‘A group of researchers at Imperial College London may have discovered a new ***cure*** *for motion sickness.’*

- semantic proximity annotation: .96

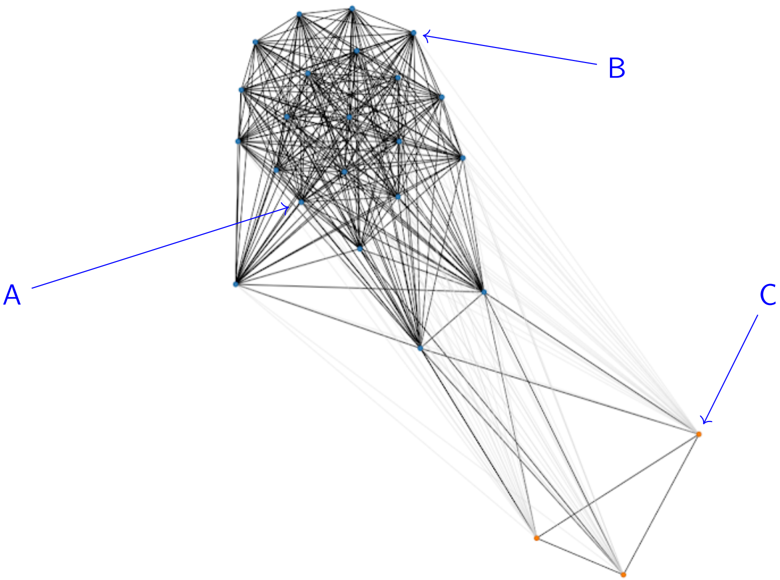
Example pair for *botemedel* ‘remedy’

(A) Medicinen stoppar RNA-virus från att föröka sig – och är snarare en bromsmedicin än ett **botemedel** .
*‘The medicine stops RNA viruses from multiplying – and is more of a brake drug than a **cure**.’*

(C) IFK Göteborg hade inget **botemedel** mot (...) Aliou Badji borta mot Djurgården .
*‘IFK Göteborg had no **remedy** against (...) Aliou Badji during their away game against Djurgården.’*

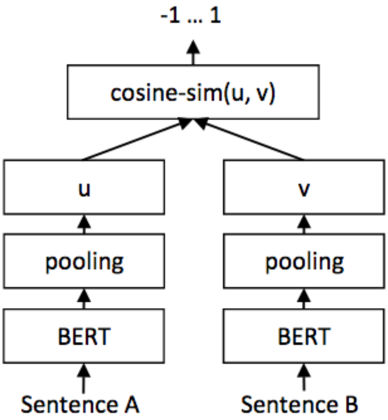
- semantic proximity annotation: .35

Example graph for *botemedel* ‘remedy’



Semantic proximity annotation

Figure 1: XL-LEXEME model architecture (Cassotti et al., 2023).



Correlation Clustering

- $w \in W$ are shifted to obtain a set of **positive** and **negative** edges
- Let $C : U \mapsto L$ be some clustering on U
- $\phi_{E,C}$ is the set of positive (high) edges **across** any of the clusters in clustering C
- $\psi_{E,C}$ the set of negative (low) edges **within** any of the clusters
- correlation clustering searches for a clustering C that minimizes the sum of weighted cluster disagreements:

$$SWD(C) = \sum_{e \in \phi_{E,C}} W(e) + \sum_{e \in \psi_{E,C}} |W(e)| \; .$$

- **main assumption**:
  - weights **above/below** threshold indicate **same/different** sense

Experiments

- **main question**: How accurate is our approach?
- multiple experimental rounds
- incremental improvements
- increasing data sizes

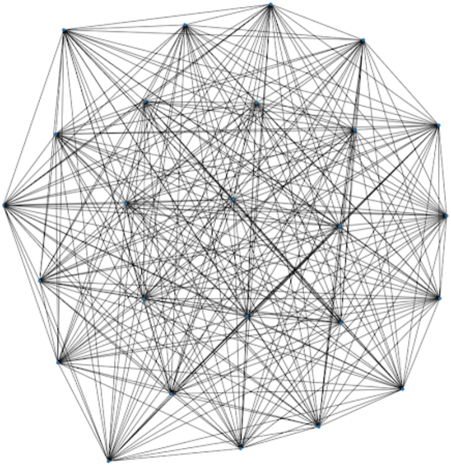
Pilot study

1. **tune** clustering parameters on 18 sense-annotated words
2. randomly select 281 **monosemous** content **words**
3. sample 25 **usages** per word
4. **cluster** with DURel tool
5. compare 1-cluster and >1-cluster groups for **non-recorded** senses

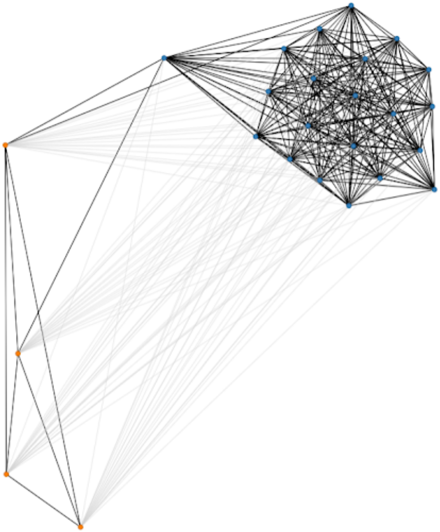
Main study

1. randomly select 1175 **monosemous** content **words**
2. sample 25 **usages** per word with improved **filtering**
3. **cluster** with DURel tool
4. compare 1-cluster and >1-cluster groups for **non-recorded** senses

Example graph for *beachvolleyboll* ‘beach volleyball’ (1-cluster)



Example graph for *skyltfönster* ‘shop window’ (>1-cluster)



Results

Table 1: Number of words with 1–6 clusters in the *pilot study* (top) and the *main study* (bottom).

	1	**2**	**3**	**4**	**5**	**6**	**total**
pilot	215	49	9	6	1	1	**281**
%	.77	.17	.03	.02	.003	.003	1.0
main	956	167	33	15	3	1	**1175**
%	.81	.14	.03	.01	.003	.003	1.0

Manual analysis

- **aim**: compare the two groups of interest (1-cluster vs. >1-cluster) for presence of **non-recorded senses**
- **expectation**: none in 1-cluster, but many in >1-cluster
- **pilot study**: randomly sample 28 words from 1-cluster group and all 66 words from >1-cluster group
- **main study**: randomly sample 21 vs. 153 words in parallel

Manual analysis

Table 2: 2×2 contingency tables for cluster number vs. non-recorded sense presence in the *pilot study* (top) and the *main study* (bottom).

Item	1-cluster	>1-cluster	Row total
non-recorded	1	30	31
recorded	27	36	63
Column total	**28**	**66**	**94**
non-recorded	0	87	87
recorded	21	66	87
Column total	**21**	**153**	**174**

Manual analysis

- >1-cluster group has a much **larger proportion** of non-recorded sense cases in both studies
- strong observed **effect size**: 4% vs. 45% (pilot), 0% vs. 57% (main)
- group differences are statistically **significant** (Fisher's exact test, $p < .01$)
- 1-cluster group is much more **frequent** than the >1-cluster group in the full data that were clustered (77% and 81% of all words)

→ 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

Dictionary updates

- A **new sense** is added to the SO database if it is considered to be **established** in modern Swedish texts (with general language rather than specialized terminology).
- **Main sense** or a **subsense**? The SO lexicographers follow the traditions applicable to the semantic description in the dictionary.

Dictionary updates

- Some examples of **headwords** in the SO database with recently added senses:
  - **main sense**:
    - *lätthet* 'lightness'; *riksmöte* 'national assembly'
  - **subsense**:
    - **figurative**: *coacha* 'coach'; *huvudrätt* 'main course', *kroppspulsåder* 'aorta'
    - **extension**: *friskförklara* 'declare healthy'; *obscen* 'obscene'
    - **generalisation**: *livförsäkring* 'life insurance'; *beroendeframkallande* 'addictive'
    - **specialization**: *drakonisk* 'draconian', *resumé* 'résumé'

Conclusion

- **simple**, **effective** approach to find non-recorded senses
- has led to several SO **updates**
- published **predictions** for ~1,500 target words[2]

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

Future work

- compare effect size across **increasing cluster numbers**
- study words with **more than one** sense in the dictionary
- incorporate **sense definitions**, and other dictionary entry information
- include **cluster size** into analysis
- incorporate more corpus **evidence**
- compare to alternative **word sense induction** models

References I

Cassotti, P., Siciliani, L., de Gemmis, M., Semeraro, G., & Basile, P. (2023, July). Xl-lexeme: Wic pretrained model for cross-lingual lexical semantic change. In *Proceedings of the 61th annual meeting of the association for computational linguistics.* Online: Association for Computational Linguistics.

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

Erk, K. (2006, June). Unknown word sense detection as outlier detection. In R. C. Moore, J. Bilmes, J. Chu-Carroll, & M. Sanderson (Eds.), *Proceedings of the human language technology conference of the NAACL, main conference* (pp. 128–135). New York City, USA: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/N06-1017`

Fedorova, M., Mickus, T., Partanen, N., Siewert, J., Spaziani, E., & Kutuzov, A. (2024, aug). AXOLOTL'24 shared task on multilingual explainable semantic change modeling. In N. Tahmasebi et al. (Eds.), *Proceedings of the 5th workshop on computational approaches to historical language change* (pp. 72–91). Bangkok, Thailand: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2024.lchange-1.8` doi: 10.18653/v1/2024.lchange-1.8

Sander, P., Hengchen, S., Zhao, W., Ma, X., Sköldberg, E., Virk, S. M., & 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.* Retrieved from `https://www.cjvt.si/wp-content/uploads/2024/10/LLM-Lex_2024_Book-of-Abstracts.pdf`

Schlechtweg, D., Virk, S. M., Sander, P., Sköldberg, E., Theuer Linke, L., Zhang, T., ... Schulte im Walde, S. (2024, mar). The DURel annotation tool: Human and computational measurement of semantic proximity, sense clusters and semantic change. In N. Aletras & O. De Clercq (Eds.), *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: System demonstrations* (pp. 137–149). St. Julians, Malta: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2024.eacl-demo.15`

Sköldberg, E., Virk, S. M., Sander, P., Hengchen, S., & Schlechtweg, D. (2024). Revealing semantic variation in Swedish using computational models of semantic proximity: Results from lexicographical experiments. In *Proceedings of the 21st EURALEX International Congress Lexicography and Semantics.* Retrieved from `https://euralex.org/publications/revealing-semantic-variation-in-swedish-using-computational-models-of-semantic-proximity-results-from-lexicographical-experiments/`