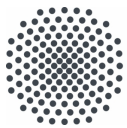
Tackling Multilingual Diachronic Unknown Sense Detection using a Few-Shot Supervised Learning Approach

DUSD

Silvia Cunico

Institut für Maschinelle Sprachverarbeitung (IMS)

Master Thesis Presentation
February, 28 2025



Universität Stuttgart

Outline

Introduction & Motivation

- **Diachronic** = Language evolves through **time** and
  1. New concepts → new words coined (*neologisms*)
  2. Existing words gain **new senses** → *polysemy* [1]

- **Unknown Sense Detection (USD)**
  1. Need to know which senses already exist for each word, a reference
  2. Unsupervised issues: polysemy goes through stages [7], need for coverage in all domains
  3. **Supervised** overcomes the issue via sense-annotated data. Issue: costly, data imbalance.

- **Motivation**: **few-shot** resource → dictionary
  1. Uniform distribution of known senses (one gloss per sense)
  2. Availability in understudied languages
  3. Structured, versioned & computer-usable [2]
  4. CHALLENGE: few representatives available per sense → this thesis.

Task Overview

- **Assess Polysemy**: dictionary as few-shot sense-inventory;
  Assess Time: asynchronous corpus w.r.t dictionary ("New")
- **Assumption**: meaning of a word in usage might NOT appear in the sense-inventory → GOAL: *known* vs. *unknown* (binary classification)

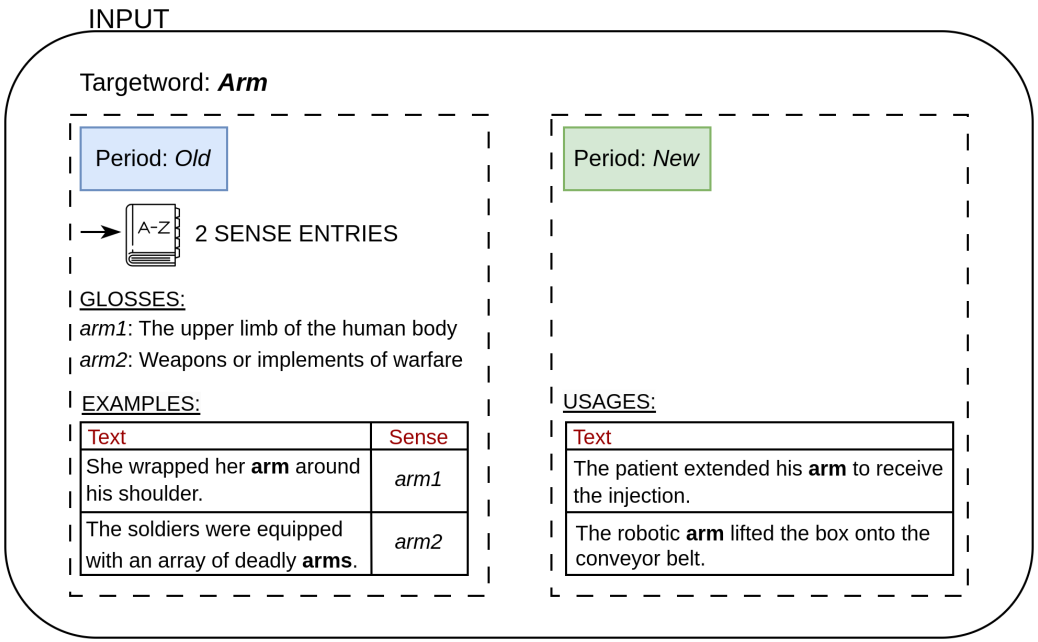


Figure: Task overview.

Task Overview

- **Assess Polysemy**: dictionary as few-shot sense-inventory;
  Assess Time: asynchronous corpus w.r.t dictionary ("New")
- **Assumption**: meaning of a word in usage might NOT appear in the sense-inventory → GOAL: *known* vs. *unknown* (binary classification)

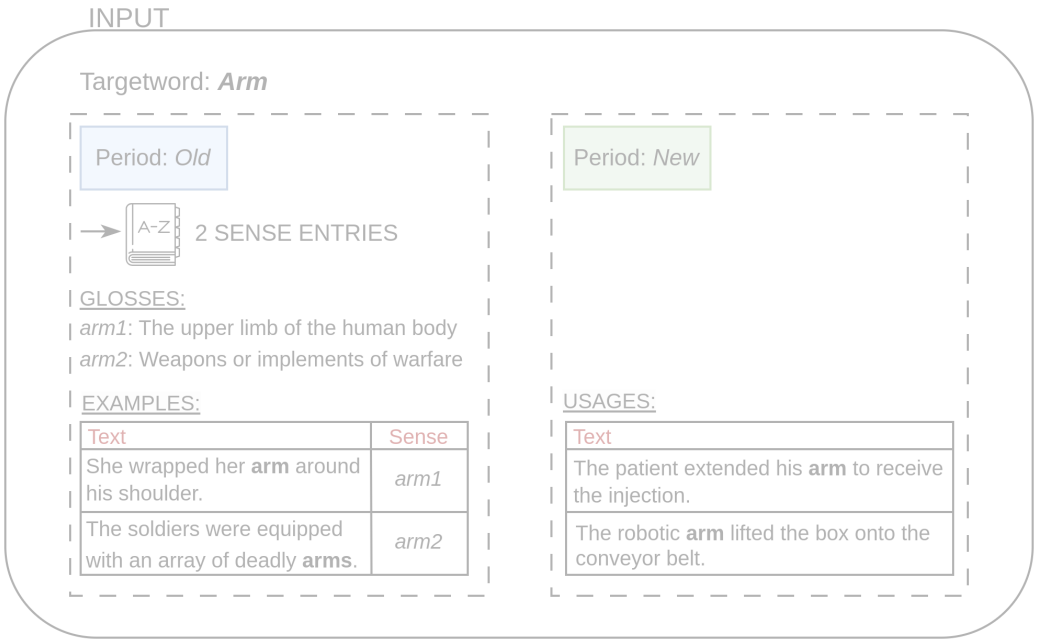


Figure: Task overview.

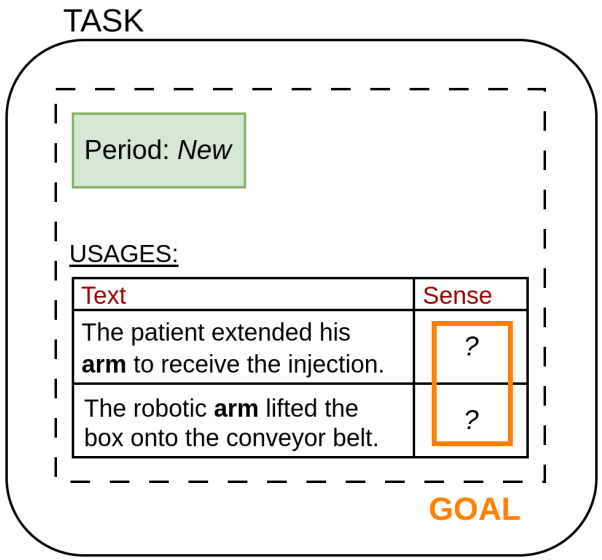


Figure: Task goal.

DUSD Task Overview

- Pave the way towards explainability of lexical semantic change for lexicographers.
- Further **GOALS:**

Subproblem	**Responsible Task**
Determining whether word usages belong to *unknown* vs. *known* sense	**Unknown Sense Detection (USD)**
Map word usages having a *known* sense with exact existing sense	**Word Sense Disambiguation (WSD)**
Group word usages with a related *unknown* sense together and assign a representative label	**Word Sense Induction (WSI)**

Related Work

1. **Unsupervised:** adopt pure WSI methods (i.e. clustering embedded word usages from "Old" vs. "New" [11], graphs overlappings [13] . . .)
2. **Supervised:**

Erk et al. [5]

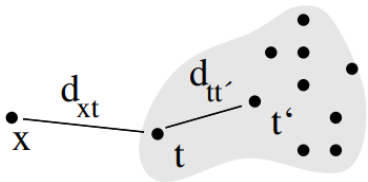


Figure: Outlier Detection approach for USD.

If $d_{xt}/d_{tt'} > \theta \rightarrow x$ outlier. Embeddings are engineered feature vectors.

Lautenschlager et al. [8]

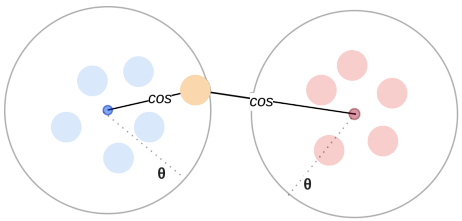


Figure: WSD-based approach for USD.

If $\max(cos, cos) < \theta \rightarrow x$ outlier. `XL-LEXEME` embeddings from augmented **glosses OR examples**.

Related Work - XL-LEXEME

WiC Pretrained Model for Cross-Lingual LEXical sEMantic changE [3]

- Assumption of temporal transferability (Cassotti et al. 2023)
- **Backbone**: `XLM-RoBERTa` (`XLM-R`) [4] (100 languages)
- **Trained on**: Word-in-Context monolingual & cross-lingual sentence pairs for 30+ languages [14, 12, 15]
- **Fine-tuning**: Sentence pairs separately encoded using a *Siamese Network*: `Sentence-BERT` (`SBERT`) [16]

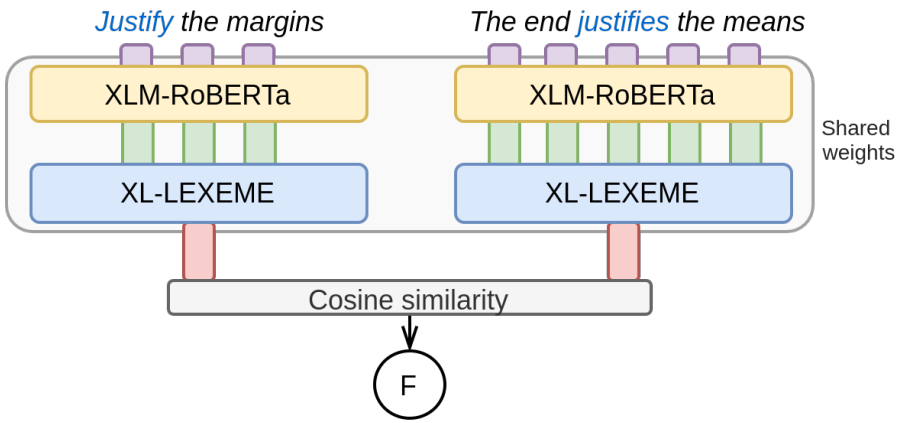


Figure: XL-LEXEME architecture.

Contributions of this thesis

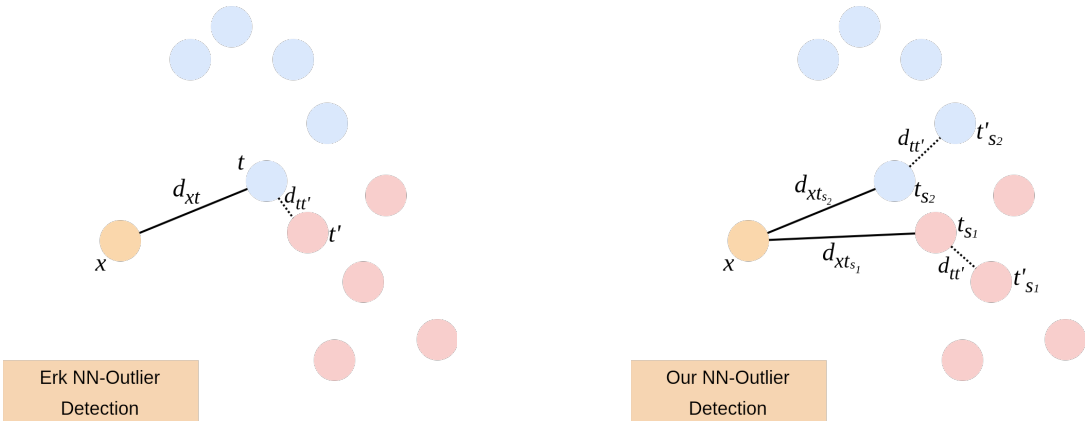
For supervised tasks USD and WSD: Sense augmentation strategies Lautenschlager et. al, from **glosses'** (3 G-augmentations) & **examples'** (2 E-augmentations) texts:

"Refers to the upper limb of the human body, extending from the shoulder to the hand" → *"Refers to the upper limb of the human body, extending from the shoulder to the hand* **i.e. arm***"*

1. M-augmentations → exploiting G and E together (M0: G1 + G2 + G3), (M1: E0 + E5), (M2: G1 + G2 + G3 + E0 + E5)
   → introducing more variations to let model recognize different surface forms of *known* vs. *unknown* sense and rely more on semantic properties rather than overfit to the sense-scarce data
2. M-augmentations on WSD-based approach with threshold but also NN-Outlier detection approach: using power of glosses (M2) despite limited information

Contributions of this thesis

3. Adaption of NN-Outlier detection approach BY SENSE:



4.
   - **Retain One** models: from the ensuing embeddings → mean or median
   - **Retain All** models: from the ensuing embeddings → outlier detection

Contribution of this thesis

5. Sense-data augmented as well as usages to classify extracted from a large-scale benchmark in more understudied languages
6. Sense-data and usage-data embedded using `XL-LEXEME` and compared by means of similarity/distance

For unsupervised WSI task:

1. Hierarchical agglomerative clustering of `XL-LEXEME` embedded usages using a threshold controlling the maximum distance at which two clusters can be merged

Data

- **AXOLOTL-24 shared task** on Multilingual Explainable Semantic Change Modeling [6]: important evaluation benchmark
- Usages from two time periods: "Old" & "New"

usage id	word	orth	sense id	gloss	example	indices	period	date
train_fi_45898	mies-luku	Mies lucu	mies-luku_XeagDiyRXpA	miesjoukko; miesten lukumäärä, pääluku	Mies lucu Bruckein ja Factoritten tykönä maalla pitä hengi lucu taxerattaman	0:9	new	1700

Figure: Extract from Finnish training dataset.

Training Data

- Languages: Finnish & Russian
- **Usage-based** sense annotations for both "Old" and "New" usages
- Fi: 93139, polysemy: 52.0%; Ru: 6494, polysemy: 79.3%

Test Data

- + surprise language: German
- OOV words w.r.t. training and development sets

Experiments Background

- **USD & WSI**: thresholds for grouping previously tuned on training datasets
- **All tasks:** **downstream tasks**, optimization through grid search → model selection

Hyperparameter	Values	Model	Task
Sense Emb. Augm. Reduction Method Usage Emb. Augm. Similarity Measure	{G1, G2, G3, E0, E5, M0, M1, M2} {Mean, Median} {ϵ, SUB} {Cosine Simil., Spearman Correl.}	Retain-One	USD/WSD
Sense Emb. Augm. Reduction Method Usage Emb. Augm. Similarity Measure	{E0, E5, M0, M1, M2} Nearest Neighbor {ϵ, SUB} {Euclidean Dist., Cosine Simil.}	Retain-All	USD/WSD
Usage Emb. Augm. Linkage Method	{ϵ, SUB} {Single-linkage, Average-linkage}		WSI

Table: Hyperparameter values used in the experiments.

Parameter	Values	Model	Task
Sim. Threshold	{0.42, 0.44, 0.46, ..., 0.98}	Retain-One	USD
Sim. Threshold	{0.7, 0.8, 0.9, ..., 1.8}	Retain-All	USD
Cluster Threshold	{0.05, 0.10, ..., 0.95}		WSI

Table: Parameter values used in the experiments.

Experiment 1 - Monolingual USD

Research Questions

- Can integrating multiple data augmentation techniques from dictionary data, leveraging both glosses for general sense descriptions and examples for specific contexts, enhance the performance of `XL-LEXEME` for USD in a few-shot setting?
- Can `XL-LEXEME` utilize a nearest-neighbor density ratio approach to improve performance on USD rather than relying on tuning similarity thresholds in a WSD-based framework?

Experiment 1 - Monolingual USD

Challenges:

- Avoid *overfitting*
- Mitigate imbalance between "New" usages with *known* vs. *unknown* sense
- Improve *generalizability* despite underrepresentativeness/sparseness

Methods:

- Erk's Masking (simulating *unknown*) [5]
- Cross-Validation (CV) stratified [1] by sense
- Threshold tuning on training datasets using CV; model selection on development datasets (OOV) embed "Old", predict "New"

Evaluation:

- Mean binary $F_{0.3}$ (decrease FPs) by word

[1] Stratified CV ensures each fold maintains the same class distribution as the full dataset, preventing bias and improving model generalization

Experiment 1 - Monolingual USD

Expectations:

- Multi augmentations boost input diversity for each sense, enhancing model robustness
- Cosine performs better (`XL-LEXEME` training)

Results:

- They do but not as significantly
- Fi ∼3% absolute improvement w.r.t to a random baseline, Ru ∼17%
- `XL-LEXEME` seems to encode Russian with more confidence
- Russian `G`-based augmentations appear to work better

(Hyper)parameter	Finnish			Russian		
	`G`	`E`	`MO`	`G`	`E`	`MO`
Sense Emb. Reducer	mean	median	mean	mean	mean	mean
Similarity Function	spearman	spearman	spearman	cosine	spearman	cosine
Usage Emb. Method	epsilon					
Similarity Threshold	0.88	0.90	0.90	0.92	0.90	0.92
Mean Binary $F_{0.3}$	0.114	0.116	**0.117**	0.686	0.668	**0.694**
Mean Binary F_1	0.134	0.135	**0.136**	0.755	0.751	**0.765**

Table: **Retain-One** models' results on test datasets.

Experiment 1 - Monolingual USD

Expectation: `M`-augmentations improve performance of *Retain-All* models, too

Results: They do but still underperform *Retain-One* models

	Model			
(Hyper)parameter	`E`	`M0`	`M1`	`M2`
Sense Emb. Reducer	erk	erk	erk	erk
Similarity Function	euclidean	euclidean	euclidean	euclidean
Usage Emb. Method	epsilon	epsilon	epsilon	epsilon
Similarity Threshold	0.70	0.80	0.70	0.70
Mean Binary $F_{0.3}$ by Word	0.018	**0.654**	0.445	**0.653**
Mean Binary F_1 by Word	0.020	**0.747**	0.513	**0.740**

Table: **Retain-All** models' results on the Russian test dataset.

Experiment 2 - Multilingual USD

Research Questions

- Can XL-LEXEME generalize to typologically diverse languages when performing USD? And how does the selection of a similarity threshold, whether optimized individually for each language or jointly across multiple languages, impact the separability of unknown senses, particularly in relation to XL-LEXEME's linguistic coverage?

Experiment 2 - Multilingual USD

Goals: Testing *cross-lingual* generalizability of `XL-LEXEME` when performing USD
Methods:

- Threshold tuning and model selection on merged datasets (Fi+Ru) → try to avoid language-specific bias
- Evaluation on 3 languages not seen during threshold tuning & model selection: German, Swedish & English [2] (data preprocessing)

Evaluation: Same as monolingual USD

[2] Courtesy of Lautenschlager et al. [8]. Data are extracted from Zenodo.

Experiment 2 - Multilingual USD

Expectation:

- Decent performance on English (*En*) & German (*De*) → extensively seen in XL-LEXEME data

Results:

- Starting language seems to make no big difference
- XL-LEXEME's only unseen Swedish (*Sv*) achieves better performance
- Notably: *Sv* seen only by backbone network XLM-R

Lang.	Augm.	*Fi*	*Ru*	*De*	*Sv*	*En*[3]
		Mean Binary $F_{0.3}$ by word				
Fi	G	0.114	0.691	0.214	0.656	0.147
	E	0.116	0.668	0.224	0.657	0.142
	M	**0.117**	**0.704**	0.224	0.657	0.147
Ru	G	**0.117**	0.686	0.210	0.656	0.147
	E	0.115	0.668	0.220	0.657	0.142
	M	0.116	0.694	0.222	0.657	0.147
FiRu	G	0.114	0.691	0.214	0.656	0.147
	E	0.116	0.663	0.219	0.657	0.142
	M	0.116	0.685	**0.231**	0.657	0.147

Table: *Retain-One* models per language on cross-lingual test sets.

Experiment 3 - DUSD IO vs. E2E

Research Questions

- Can a stacked architecture integrating USD, WSD, and WSI subtasks in a downstream manner achieve state-of-the-art performance in diachronic unknown sense detection without fine-tuning the sense encoder weights? Furthermore, how does the joint optimization of these tasks compare to their independent tuning in terms of overall performance?

Experiment 3 - DUSD IO vs. E2E

Goal: Full DUSD with disambiguation component for both *known* & *unknown*. Comparison of our method to other SOTA approaches for diachronic USD.
Methods: First, tasks individually optimized (IO), prediction using best performing models (on dev). Secondly, joint optimization of the tasks (E2E).
Evaluation:

- Macro-F_1 by word:
  1. Averages F_1 score by word
  2. Only for "New" usages with *known* senses
- Adjusted Random Index (ARI):
  1. Similarity of pairs of clusters
  2. For "New" usages with both *known* AND *unknown* senses

Experiment 3 - DUSD IO vs. E2E

Expectations: If the threshold is not too dependent on USD then overall results improve.

Results:

- E2E results ensue indeed from lower USD thresholds (lowest!): 0.35
- Boosted performance then given by more powerful WSD task

		DUSD-IO	DUSD-E2E
Fi:	F_1:	0.131	**0.714**
	ARI:	0.418	**0.576**
Ru:	F_1:	0.249	**0.702**
	ARI:	0.063	**0.113**

Table: Results on the development sets for DUSD-IO vs. DUSD-E2E.

Experiment 3 - DUSD IO vs. E2E

Comparison with other AXOLOTL-24 teams:

Team	Fi		Ru		De	
	F_1	ARI	F_1	ARI	F_1	ARI
Baseline	0.230	0.023	0.260	0.079	0.130	0.022
Deep-Change	**0.756**	**0.638**	**0.750**	0.059[4]	**0.758**	**0.543**
Holotniekat	0.655	0.596◇	0.661	0.043	0.608	0.298
ABDN-NLP	0.590	0.553	0.570	0.009	0.638	0.102
TartuNLP	0.550	0.437	0.640	0.098	0.580	0.396
WooperNLP	0.503	0.428	0.446	0.132◇	0.000	0.000
DUSD-E2E (Us)	0.686◇	0.580	0.719◇	0.115◇	0.752◇	0.412◇

Table: Performance comparison across teams.

Take-Home Messages

- M-augmentations can improve the performance on semantic-related downstream tasks in few-shot scenario
- M-augmentations can enable an outlier detection approach to perform USD
- Our models are not able to generalize to seen languages like English and German but are able to generalize to unseen language Swedish when performing USD
- We achieve competitive performance in AXOLOTL-24 without fine-tuning XL-LEXEME
- FUTURE WORK: explore more augmentations, longer time span between lexical resource and corpus of usages to classify, language-conditioned thresholds

Thank you for listening!

Your questions and insights are most welcome.

Should time be insufficient, you may reach me at:
st179785@stud.uni-stuttgart.de

References I

[1] Michel Bréal. *Essai de sémantique. Science des significations*. Paris: Hachette, 1897.

[2] Bruce G. Buchanan and David C. Wilkins, eds. *Readings in knowledge acquisition and learning: automating the construction and improvement of expert systems*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN: 1558601635.

[3] Pierluigi Cassotti et al. "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)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 1577–1585. DOI: 10.18653/v1/2023.acl-short.135. URL: https://aclanthology.org/2023.acl-short.135/.

References II

[4] Alexis Conneau et al. "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://aclanthology.org/2020.acl-main.747/.

[5] Katrin Erk. "Unknown word sense detection as outlier detection". In: *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. Ed. by Robert C. Moore et al. New York City, USA: Association for Computational Linguistics, June 2006, pp. 128–135. URL: https://aclanthology.org/N06-1017.

[6] Mariia Fedorova et al. *AXOLOTL'24 Shared Task on Multilingual Explainable Semantic Change Modeling*. 2024. arXiv: 2407.04079 [cs.CL]. URL: https://arxiv.org/abs/2407.04079.

[7] Paul J. Hopper et al. "On Some Principles of Grammaticization". In: *Approaches to Grammaticalization* 1 (1991), pp. 17–35.

References III

[8] Jonathan Lautenschlager et al. *Detection of Non-recorded Word Senses in English and Swedish*. 2024. arXiv: 2403.02285 [cs.CL]. URL: https://arxiv.org/abs/2403.02285.

[9] Fuli Luo et al. "Incorporating Glosses into Neural Word Sense Disambiguation". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 2473–2482. DOI: 10.18653/v1/P18-1230. URL: https://aclanthology.org/P18-1230/.

[10] Fuli Luo et al. "Leveraging Gloss Knowledge in Neural Word Sense Disambiguation by Hierarchical Co-Attention". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 1402–1411. DOI: 10.18653/v1/D18-1170. URL: https://aclanthology.org/D18-1170/.

References IV

[11] Xianghe Ma, Michael Strube, and Wei Zhao. "Graph-based Clustering for Detecting Semantic Change Across Time and Languages". In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Yvette Graham and Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 1542–1561. URL: https://aclanthology.org/2024.eacl-long.93/.

[12] Federico Martelli et al. "SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC)". In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Ed. by Alexis Palmer et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 24–36. DOI: 10.18653/v1/2021.semeval-1.3. URL: https://aclanthology.org/2021.semeval-1.3/.

References V

[13] Sunny Mitra et al. "An automatic approach to identify word sense changes in text media across timescales". In: *Natural Language Engineering* 21.5 (2015), pp. 773–798. DOI: 10.1017/S135132491500011X.

[14] Mohammad Taher Pilehvar and Jose Camacho-Collados. "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)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 1267–1273. DOI: 10.18653/v1/N19-1128. URL: https://aclanthology.org/N19-1128.

References VI

[15] Alessandro Raganato et al. "XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 7193–7206. DOI: 10.18653/v1/2020.emnlp-main.584. URL: https://aclanthology.org/2020.emnlp-main.584/.

[16] Nils Reimers and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. 2019. arXiv: 1908.10084 [cs.CL]. URL: https://arxiv.org/abs/1908.10084.

Extra - Masking & CV

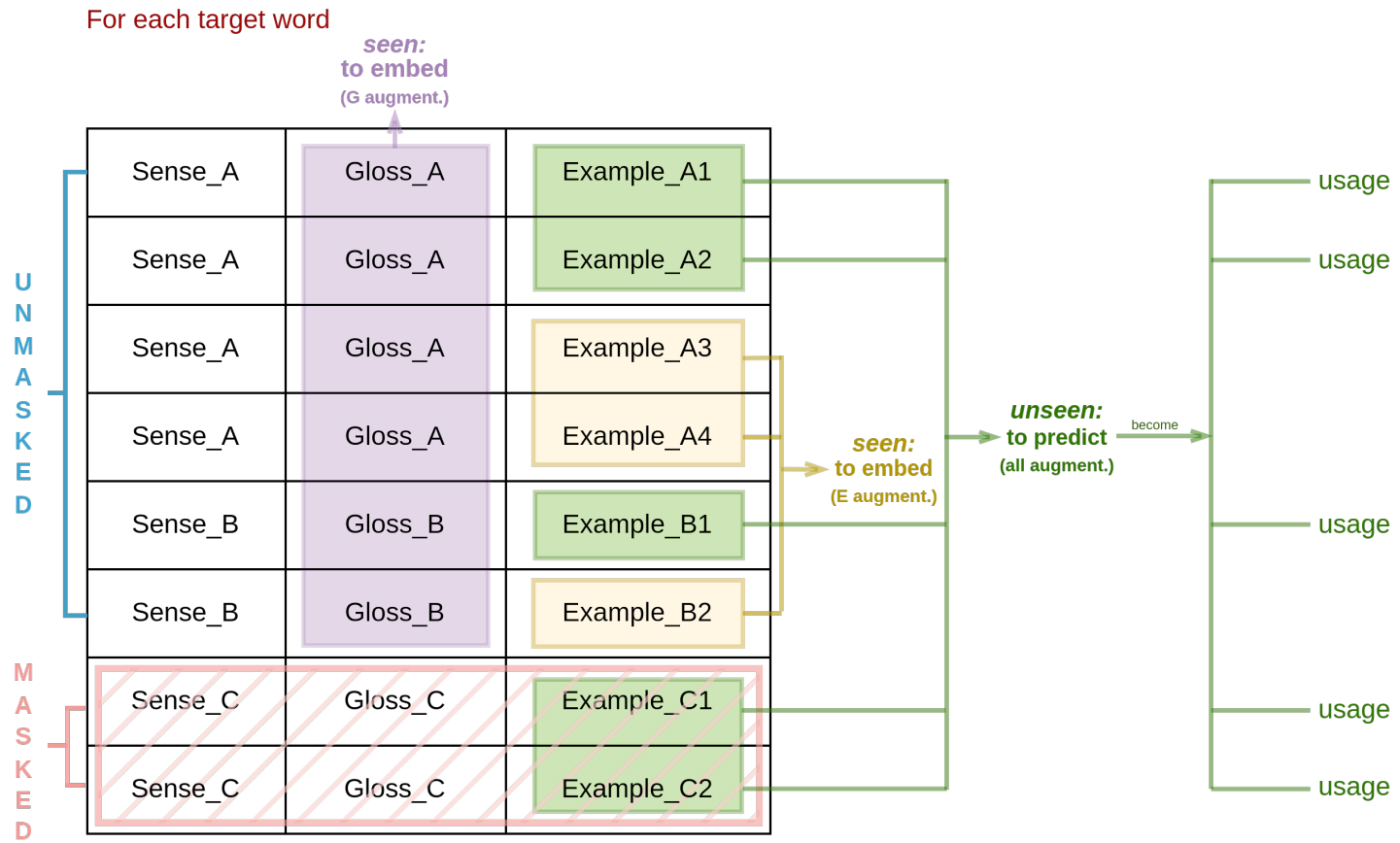


Figure: Masking and CV.

Motivation Stratification by Sense

Why I chose CV by sense: Let's consider the Finnish word *aivan* (extracted from the given AXOLOTL-24 Finnish training set), which has three senses:

- **Sense 1:** *(us. taipumaton): pelkkä, paljas, sula; yksinomainen*
  Meaning: "mere, bare, pure; exclusive"
  Number of examples: **6**
- **Sense 2:** *ihan; vallan, suorastaan, peräti, kovin*
  Meaning: "quite, entirely, absolutely, downright, very"
  Number of examples: **26**
- **Sense 3:** *pelkästään*
  Meaning: "only, solely"
  Number of examples: **2**

By using the by sense stratification, I make sure that the least frequent sense (*pelkästään*, with only 2 examples) is present in as many evaluation folds as possible. In this way I wish to ensure a more balanced evaluation.

Extra - Frequency of *known* vs. *unknown*

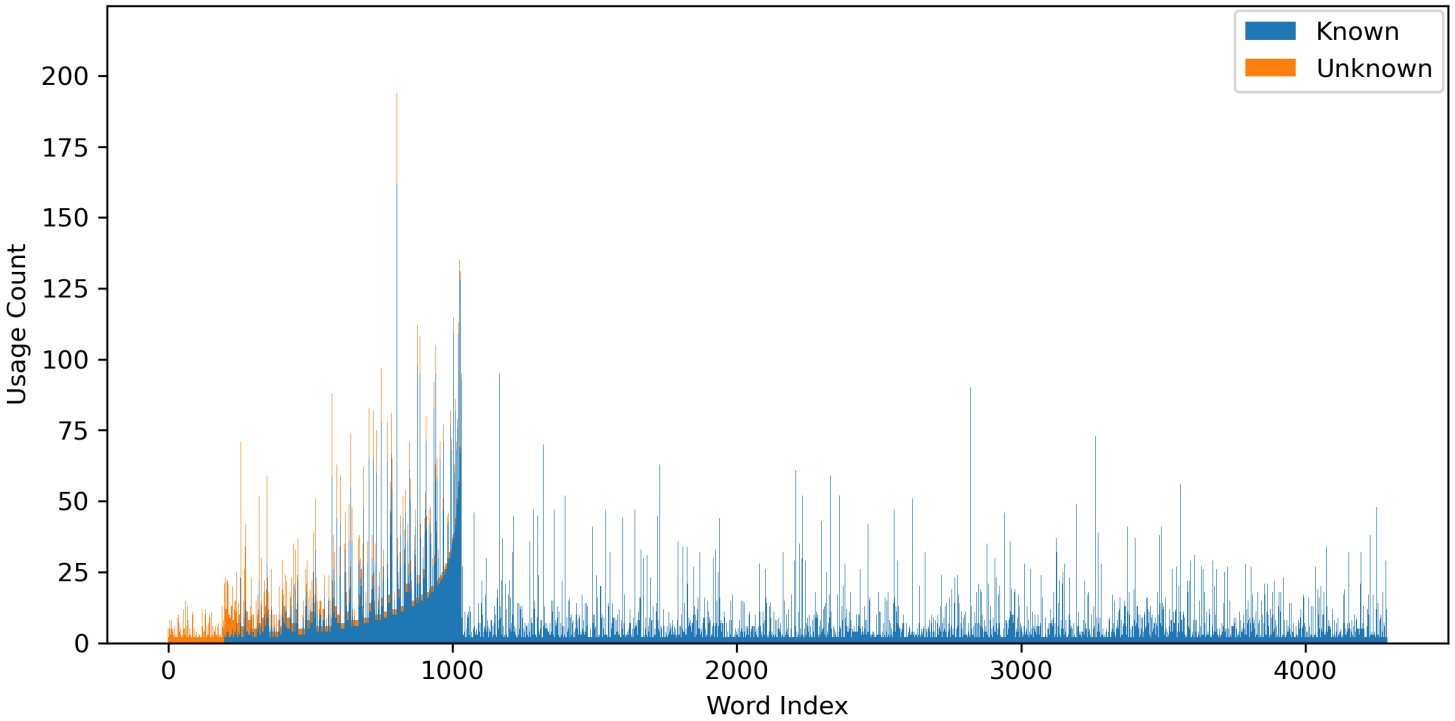


Figure: Frequency of known vs. unknown usages per word index in the Finnish training set.

Extra - Cluster Agreement

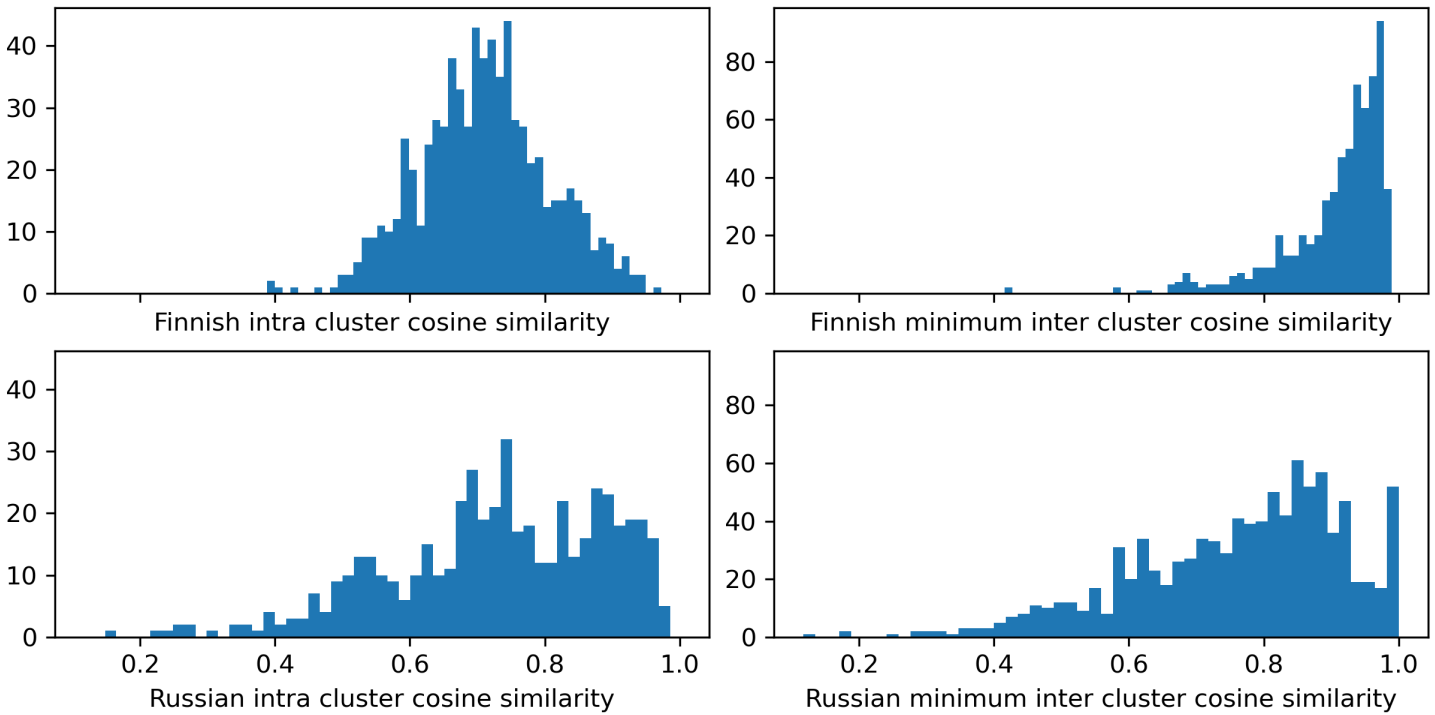


Figure: Comparison of Cluster Pairwise Similarity of Finnish (above) vs. Russian (be- low).