Detecting Subtle Sense Shift with Polysemy-Aware Trends

Ondřej Herman[1,2], Pavel Rychlý[1,2]

{xherman1,pary}@fi.muni.cz

[1]Natural Language Processing Centre, Masaryk University, Brno, Czech Republic
[2]Lexical Computing, Brno, Czech Republic

Abstract

Language changes faster than dictionaries can be revised, yet automatic tools still struggle to spot the subtle, short-term shifts in meaning that precede a formal update. We present a language-independent pipeline that detects word-sense shifts in large, time-stamped web corpora. The method couples a robust re-implementation of the Adaptive Skip-Gram model, which induces multiple sense vectors per lemma without any external inventory, with a second stage that tracks each sense through time under three alternative frequency normalizations. Linear Regression and the robust Mann-Kendall/Theil–Sen estimator then test whether a sense's frequency slope deviates significantly from zero, producing a ranked list of headwords whose semantics are drifting.

We evaluate the system on the English (12 B tokens) and Czech (1 B tokens) Timestamped corpora for May 2023–May 2025. Expert annotation of the top-100 candidates for each model variant shows that 50.7% of Czech and 25.7% of English headwords exhibit genuine sense shifts, despite web-scale noise.

1 Introduction

Dictionaries lag behind the language they purport to describe: while web discourse spreads new shades of meaning within weeks, lexicographic updates arrive months or years later. Automating the early detection of *word-sense shifts* would help editors allocate effort to the most dynamic headwords and, in turn, keep lexical resources current.

Existing research on lexical semantic change focuses on century-scale, well-curated corpora and binary "changed / unchanged" judgements (Schlechtweg, 2023; Tahmasebi and Dubossarsky, 2023). We argue that today's monitor corpora—billions of time-stamped web tokens—call for a different objective: spotting *emerging or drifting senses* in near-real time, despite noise and domain volatility.

We propose a three-stage, language-independent pipeline for tracking within-lemma sense redistribution that

1. induces polysemy-aware embeddings with an efficient **Adaptive Skip-Gram** implementation (§4.2),
2. tracks each induced sense through time and tests for statistically significant trends (§4.4),
3. outputs ranked headwords that lexicographers can inspect (§6).

Evaluated on **12 B English** and **1 B Czech** tokens (May 2023–May 2025), the method delivers a **25–51%** precision of genuinely new or rising senses, according to expert annotation.

2 Related Work

Lexical Semantic Change Detection (LSCD). Early LSCD studies relied on co-occurrence vectors (Sagi et al., 2009; Gulordava and Baroni, 2011) or topic models (Cook et al., 2013). Recent competitions (Schlechtweg et al., 2020) popularised embedding-distance measures, yet still treat time as two coarse epochs.

Sense-aware embedding models. Multi-prototype embeddings such as Adaptive Skip-Gram (Adagram) (Bartunov et al., 2016a) and later transformer-based approaches capture polysemy more explicitly than single-vector models. However, few works couple them with diachronic trend analysis.

Trend detection in corpora. Herman and Kovar (2013) introduced frequency normalization and statistical trend estimation for neologism detection. We extend this idea to sense-level frequencies.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 60–65

Corpus	Tokens	Time span	Docs/day
English	12.0 B	2023-05 – 2025-05	21.1 k
Czech	1.1 B	2023-05 – 2025-05	2.6 k

Table 1: Sketch Engine Monitor corpora used in this study.

Lexicographic applications. Commercial platforms (e.g. Sketch Engine; Kilgarriff et al., 2014) offer keyword-in-context views but no automatic sense-shift alerts. Our pipeline is designed to plug directly into such workflows.

3 Data: Timestamped Web Corpora

We work with the English and Czech Timestamped monitor corpora (Herman et al., 2025) distributed in Sketch Engine.[1] The corpus construction pipeline is described in Krek et al. (2017). Each document comes with an RSS publication timestamp. The text is extracted from HTML using jusText (Pomikálek, 2011), tokenised by Unitok (Michelfeit et al., 2014), and deduplicated on paragraph level using Onion (Pomikálek, 2011), and POS-tagged + lemmatised using TreeTagger (Schmid, 2013).

Their size and daily growth make them ideal test-beds for short-term semantic drift.

4 Methodology

Our pipeline decouples *sense induction* (§4.2) from *trend estimation* (§4.4), each oblivious to language specific issues. To identify trending senses, we build a word sense induction model based on the source corpus, and extract the senses of the examined headwords. We then calculate the frequencies of the senses over time, estimate the statistical trend, and rank the results.

4.1 Word Sense Induction

We use Adaptive Skip-gram (Bartunov et al., 2016b) to induce word senses from the corpus text. This algorithm satisfies multiple practical constraints that arise when tracking short-term meaning change in web-scale monitor corpora:

Language and language-resource independence. It learns multiple prototypes for a word directly from raw context, avoiding reliance on external sense inventories that either do not exist (for low-resource languages) or lag behind emerging usage. This capability is essential for our goal of detecting senses that were *unknown at training time* – a scenario where large language models fall short, since their language knowledge is frozen at the moment of pre-training.

Computationally efficient. The training loop mirrors classic skip-gram, with only a constant-factor overhead for sense selection, enabling frequent retraining.

Representation quality. Intrinsic evaluation on the ShadowSense benchmark (Herman and Jakubíček, 2024) confirms that Adaptive skip-gram's induced clusters align with human intuition for a majority of test words, providing the semantic fidelity required for downstream trend analysis. While more recent transformer-based multi-prototype models achieve marginal gains on static datasets, their heavier resource demands and dependence on pre-training make them ill-suited for the continuously growing corpora that motivate our study. In short, Adagram delivers the right mix of quality, scalability, and resource independence needed to monitor lexical change in real time.

4.2 Sense Induction with Adaptive Skip-Gram

To obtain the word sense induction model, we train Adaptive skip-gram on raw tokens, allowing up to $K=10$ senses per headword. To reach web scale, we re-implemented the original Julia code in Rust, leveraging compact Manatee (Rychly, 2007) corpus representation for efficient access. The implementation is available at `https://github.com/ondra/sensetrends`. Computing 64-dimensional embeddings over one training pass on 12 B tokens takes 60 hours using 32 threads on AMD EPYC 9654 CPU.[2]

Each token is then assigned to its most probable sense and tabulated by the diachronic epoch.

4.3 Diachronic Frequency Normalization

At this point, we have obtained the frequency distributions of the senses over time for each word, which are not directly comparable. We explore three normalization approaches, each highlighting a different aspect of change.

For every headword, S is the set of senses and E is the set of the epochs, (e.g. months) present

[1] Accessible through the interface at `https://www.sketchengine.eu` and made for linguistic research. The full corpus text is available for research purposes .

[2] Using more threads does not speed up the computation significantly due to memory access contention.

in the corpus. Let $f_{\text{raw}}(s, e)$ be the count of sense $s \in S$ in epoch $e \in E$, $N(e)$ the token total of e.

Epoch-normalized

$$f_{en}(s, e) = \frac{|E|\ f_{\text{raw}}(s, e)/N(e)}{\sum_{e'} f_{\text{raw}}(s, e')/N(e')}$$

A natural extension of the approach used by Herman and Kovar (2013) for the analysis of the diachronic behavior of words. Every sense is examined in isolation and sums to $|E|$, so for a perfectly stable sense, f_{en} will be 1 everywhere. Peaks indicate epochs where s is over-represented relative to its own history.

Global-normalized

$$f_{gn}(s, e) = \frac{|S||E|\ f_{\text{raw}}(s, e)/N(e)}{\sum_{s'} \sum_{e'} f_{\text{raw}}(s', e')/N(e')}$$

Similar to f_{en}, but reserves proportionality across the senses based on the corpus frequency. The dominance of rare senses is reduced compared to f_{en}, which treats every sense in isolation.

Sense-relative

$$f_{sr}(s, e) = \frac{f_{\text{raw}}(s, e)}{\sum_{s'} f_{\text{raw}}(s', e)}$$

Captures *within-word* redistribution of meaning. The f_{sr} sums to 1 over all senses within every epoch. This approach removes the dependence on absolute corpus frequencies and correlates with sense shift closely.

4.4 Trend Estimation

In this step, we estimate the rate of change and the corresponding p-value for every sense.

Given the time series $\{(x_i, y_i)\}$, where x_i indexes epochs and y_i is the raw corpus frequency normalized by f_{en}, f_{gn}, or f_{sr}, we fit (i) ordinary least squares (OLS) and (ii) the Theil–Sen (TS) robust slope estimator accompanied by the Mann-Kendall statistical test as described in Herman and Kovar (2013). In our experiment, a sense is flagged if $p < 10^{-4}$.

4.5 Candidate Ranking

For each headword, we keep the steepest trending sense, then sort words by the slope magnitude. The pipeline outputs top-k words for review.

5 Experimental Setup

Vocabulary. We analyze the diachronic behavior of the 30,000 most frequent lemmas, excluding those starting with initial capital as a rudimentary form of named entity recognition.

Epoch size. Monthly windows offer a compromise between stability and responsiveness (≈ 800 M English tokens per diachronic epoch).

Annotation protocol. A trained in-house linguist labeled the top 100 candidates sorted according to the trend slope from each of six method variants (3 norms and 2 slope estimators) as

1. OK - potentially of lexicographical interest
2. BAD - no lexicographical relevance / senses not understandable
3. ERROR - other processing error / bad headword

The total number of headwords was approximately 250 for each of the languages due to overlap between the lists.

The annotator was shown the frequency plot for every sense, with the senses described by a list of their nearest neighbors in the word embedding space of the Adagram model, as seen in Figure 1.

6 Evaluation and Results

Norm	Est.	English (12 B)		
		OK	Bad	Err
f_{en}	OLS	28	57	15
f_{en}	TS	24	64	12
f_{gn}	OLS	26	59	15
f_{gn}	TS	24	62	14
f_{sr}	OLS	24	58	18
f_{sr}	TS	22	62	16

Table 2: English precision at 100.

Norm	Est.	Czech (1.1 B)		
		OK	Bad	Err
f_{en}	OLS	49	47	4
f_{en}	TS	50	49	1
f_{gn}	OLS	49	47	4
f_{gn}	TS	50	48	2
f_{sr}	OLS	52	46	2
f_{sr}	TS	54	43	3

Table 3: Czech precision at 100.

Findings. The Czech corpus attains a mean 50.7% hit-rate, twice the English 25.7%. We attribute the gap to heavier spam and boiler-plate contamination in English web feeds, and larger absolute corpus size, which amplifies small, but still statistically significant artifacts.

6.1 Sample results

6.1.1 Annotated as *OK*

Figure 1 plots weekly sense shares for the English noun *rag*; a new AI-related "retrieval-augmented generation" sense appears at the end of 2024. Additional interesting examples include the English noun *whale* (animal → influential investor), and Czech *motorista* (motorist → member of a political party), or *box* (trending sense describing a self-pickup delivery box).

6.1.2 Annotated as *Bad*

The annotator understood the word, but the change was not meaningful, interesting, or the trending sense was not possible to distinguish from the others. For example, *fertilizer* trending within stock news, or *integration* trending in contexts related to market research, where the sense carried by the word is not truly different, but separated by the WSI algorithm due to its specificity, or *signup* trending within fantasy footbal contexts, which were not present in earlier versions of the underlying corpus.

6.1.3 Annotated as *Error*

Typical representatives of tokens marked as *error* are tokens, which were not recognized as words in the target language by the annotator, e.g. *"6a"* trending in figure labels, *+1* as a part of a phone number, ˆ (a single caret character) trending within stock report contexts, or URLs or their parts.

7 Discussion

The expert evaluation found a **50.7% precision** for Czech versus **25.7% for English**. While this might seem low, the evidence-based alternative is the direct analysis of the whole candidate list, so our method has the potential to save significant amount of work.

7.1 Czech and English Result Difference

Corpus composition. The English Timestamped corpus is almost an order of magnitude larger. More sources imply more random shift bursts and boiler-plate fragments that survive text cleaning and deduplication and inflate sense counts. English is an important language for our research, so we also tend to improve the processing and add new data sources more often compared to Czech, introducing shifts in frequencies.

Spam vulnerability. English is a primary target for SEO and click-bait content that is hard to identify at crawl time; such noise produces statistically significant but linguistically meaningless trends. Currently, websites appear to be all or nothing with respect to spam content, but the extracted text is locally plausible – no boilerplate remains, and the issues only emerge in aggregate. LLMs in the hands of spammers significantly complicate the direct detection of spam text, so this issue will only get worse.

Relative domain stability. Czech web feeds are fewer and more homogeneous, reducing abrupt register or genre shifts that can masquerade as semantic change.

7.2 What about recall?

Unfortunately, we are only able to report precision and not recall at this time, as we do not have any grounded data describing word sense shift along with a compatible corpus source. As a proxy for estimating recall, we carried out experiments with sampling headwords, for which no significant trend was identified, but we failed to find any false negatives in a sample of 500 lemmas. In the future, we intend to compare our result with a revision of a dictionary to obtain a more objective insight into the quality of the methods.

7.3 Three Normalization Strategies

The result lists for the three normalization strategies turned out to be very similar with a large overlap, contrary to our expectations. Even though the ordering of the result is not significantly different, the estimates have different and meaningful interpretations for the direct quantification of the change. Trend estimate of f_{en} represents the diachronic change of a single sense in isolation; f_{gn} introduces weighting by sense frequency within a single headword, and f_{sr} quantifies the redistribution between different senses.

8 Future Work

Evaluation. The current evaluation is quite limited, so we plan to explore the results for other

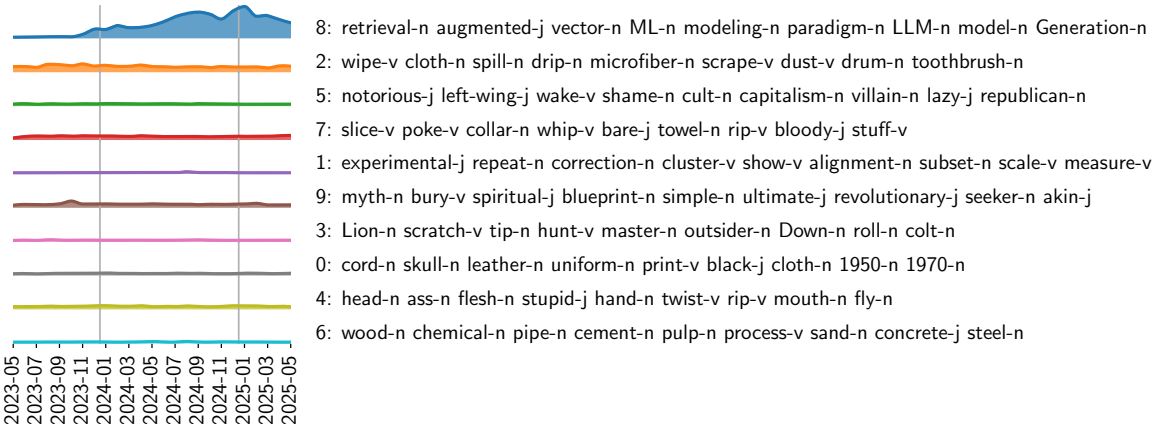

Figure 1: Emerging AI sense of *rag*.

languages and for larger result lists. We also intend to measure inter-annotator agreement.

Artifact cleanup. The method is sensitive to changes in corpus composition shifts and to all kinds of processing changes and issues, such as encoding and boilerplate removal errors. These issues tend to be strongly associated with specific feeds, so we want to integrate feed-level weighting to these types of artifacts.

Improve sense descriptions. Annotators found nearest neighbors cryptic or opaque. We believe tightly bound collocations as disambiguators should provide superior results.

9 Conclusion

We presented a pipeline that combines polysemy-aware embeddings with robust trend statistics to identify short-term word-sense shifts in billion-token web corpora. The method achieves up to 54% precision in Czech and offers immediately actionable headword queues for lexicographers.

Limitations

Annotation coverage. Only one annotator judged the sample. Inter-annotator agreement remains to be measured.

Language coverage. We tested the results for two Indo-European languages. Other language families or low-resource languages may need different preprocessing.

Bias in web sources. Monitor corpora skew towards news outlets, potentially missing sense change in spoken or social media registers.

New meaning or just different context? It is sometimes difficult to differentiate between a distinct sense of a word and a specific context a word appears in. WSI is a hard problem and remains unsolved (Mosolova et al., 2025). Even agreement between humans is often low (Kilgarriff, 1997; Herman and Jakubíček, 2024).

Acknowledgments

We used an AI assistant to improve the language of the article.

The work described herein has also been using tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062)

The described study comes from the project "On our own: Opportunities and Risks in the Individualization of Society (PRINS) CZ.02.01.01/00/23_025/0008710", which is co-financed by the European Union.

References

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2016a. Breaking sticks and ambiguities with adaptive skip-gram. In *artificial intelligence and statistics*, pages 130–138. PMLR.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2016b. Breaking sticks and ambiguities with adaptive skip-gram. In *Artificial Intelligence and Statistics*, pages 130–138.

Paul Cook, Jey Han Lau, Michael Rundell, Diana McCarthy, and Timothy Baldwin. 2013. A lexicographic appraisal of an automatic approach for detecting new word senses. *Proceedings of eLex*, pages 49–65.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic

change in the google books ngram corpus. In *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, pages 67–71.

O. Herman, M. Jakubíček, J. Kraus, and V. Suchomel. 2025. From word of the year to word of the week: Daily-updated monitor corpora for 25 languages. *Electronic lexicography in the 21st century. Proceedings of the eLex 2025 conference*.

Ondřej Herman and Miloš Jakubíček. 2024. Shadowsense: a multi-annotated dataset for evaluating word sense induction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14763–14769.

Ondrej Herman and Vojtech Kovar. 2013. Methods for detection of word usage over time. In *Seventh Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2013*, pages 79–85, Brno. Tribun EU.

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Adam Kilgarriff, Vit Baisa, Jan Busta, Milos Jakubicek, Vojtech Kovar, Jan Michelfeit, Pavel Rychly, and Vit Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

Simon Krek, Ondřej Herman, Jan Bušta, Miloš Jakubíček, and Blaž Novak. 2017. Jsi newsfeed corpus. In *The 9th International Corpus Linguistics Conference*. University of Birmingham.

Jan Michelfeit, Jan Pomikálek, and Vít Suchomel. 2014. Text tokenisation using unitok. In *RASLAN 2014*, pages 71–75, Brno, Czech Republic. Tribun EU.

Anna Mosolova, Marie Candito, and Carlos Ramisch. 2025. In the LLM era, word sense induction remains unsolved. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17161–17178, Vienna, Austria. Association for Computational Linguistics.

Jan Pomikálek. 2011. *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic.

Pavel Rychly. 2007. Manatee/bonito–a modular corpus manager. *RASLAN 2007 Recent Advances in Slavonic Natural Language Processing*, page 65.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111. Association for Computational Linguistics.

Dominik Schlechtweg. 2023. *Human and computational measurement of lexical semantic change*. Ph.D. thesis, Universität Stuttgart.

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

Helmut Schmid. 2013. Probabilistic part-of-speech tagging using decision trees. In *New methods in language processing*, pages 154–164. Routledge.

Nina Tahmasebi and Haim Dubossarsky. 2023. Computational modeling of semantic change. *Preprint*, arXiv:2304.06337.