Evaluating Trends in Czech and English

Ondřej Herman

Natural Language Processing Centre, Masaryk University
Brno, Czech Republic
xherman1@fi.muni.cz
http://nlp.fi.muni.cz

Abstract. Automated trend detection is a crucial tool for monitoring language evolution in large diachronic corpora, but its raw output mixes salient neologisms with statistical noise. This paper presents a systematic evaluation of Sketch Engine's trends feature, analyzing the top 100 trending lemmas from the Czech and English Trends corpora across one-year (2024-2025) and two-year (2023-2025) periods. Using a six-category classification scheme, a human expert annotated the lemmas to distinguish genuinely interesting trends from artifacts like proper names, processing errors, and seasonal usage. To assess the feasibility of automating this qualitative analysis, we benchmarked the human judgments against those of the GPT-OSS 120B large language model on the same task.
Our findings show that between 47% and 59% of the candidates identified by the human were valid trends. The analysis also reveals that longer, two-year timeframes effectively filter out seasonal words that are flagged in a one-year period. Human-LLM agreement varied from fair to substantial (Cohen's $\kappa = 0.288 - 0.602$), with the model systematically overidentifying 'source composition issues' and 'seasonal words' compared to the expert. The study provides a quantitative baseline for the quality of trend detection output and offers insights into the current capabilities and limitations of LLMs for sophisticated data filtering in corpus linguistics.

Keywords: Corpus linguistics; neology; trends.

1 Introduction

The rapid evolution of language, driven by digital communication and global events, presents a constant challenge for lexicographers, linguists, and developers of natural language processing tools. While large, continuously updated monitor corpora provide the necessary data to track these changes, automated methods are required to sift through billions of tokens to identify meaningful lexical trends.

This paper presents a systematic evaluation of the trends feature within the Sketch Engine corpus analysis platform. The feature is designed to detect words with a statistically significant increase in frequency over time. However, the raw output of such a system often includes not only linguistically interesting neologisms but also artifacts such as proper names, seasonal terms, data processing errors, and phenomena related to shifts in the corpus source composition.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2025, pp. 187–194, 2025.

This article analyzes the top 100 trending lemmas from the Czech Trends and English Trends corpora across two timeframes: a recent one-year period (October 2024-September 2025) and a two-year period (February 2023-February 2025). A human annotator classified each trending lemma into one of six categories to distinguish between genuinely interesting trends and various types of noise. We also explore the potential for automating this evaluation, comparing human judgments with the classification decisions of the GPT-OSS 120B LLM. Our results provide insights into the quality of the practical output of automated trend detection systems and assess the capability of LLMs to perform this type of qualitative linguistic analysis.

2 Corpus Data

The experiments were conducted on the English Trends and Czech Trends corpora, which are large scale diachronic monitor corpora available within the Sketch Engine platform [2]. These corpora are designed for tracking recent language changes and are continuously updated with new text data.

Data Source and Construction. The corpora are built from text content gathered from the web via RSS and Atom news feeds. This collection method provides reliable, finegrained publication timestamps for each document, which is a prerequisite for robust trend analysis. The crawling process began around 2014 at the Jožef Stefan Institute [8] and has been expanded since 2021 with a custom-built web feed crawler [1]. Each document in the corpus is annotated with metadata, including its publication date, source URL, and feed URL.

Text Processing Pipeline. To ensure data quality and consistency, all texts are processed using a standardized processing pipeline. First, meaningful content is extracted from the raw HTML of web pages using the `JusText` tool, which removes boilerplate content such as headers, footers, and navigation menus [4]. Subsequently, the `Onion` [4] tool is used to identify and remove duplicate and near-duplicate paragraphs from the collected texts.

The cleaned text is then linguistically annotated. Tokenization is performed using `Unitok` [3], followed by part-of-speech tagging and lemmatization with the `TreeTagger` [6] for the English Trends corpus, and Majka [7] for the Czech Trends corpus. Finally, the processed and annotated text is indexed using the `manatee` corpus manager to provide efficient querying and analysis of the data [5]. The continuous ingestion of timestamped data and this robust processing pipeline make these corpora a practical resource for investigating contemporary lexical trends.

3 Methodology and Experimental Setup

The evaluation is designed to quantify the practical quality of an automated trend detection system and to compare human judgment against that of a large language model for the task of classifying or filtering its output.

3.1 Trend Detection

The initial set of candidate words was generated using the Trends feature of the Sketch Engine corpus manager, which identifies words whose usage frequency shows a statistically significant increase over a specified time period. We used the robust statistics based method, which combines two non-parametric statistical techniques:

- The **Mann-Kendall test** is used to determine if a monotonic trend (either increasing or decreasing) exists in the time-series data of a word's frequency, providing a p-value, which describes the level of statistical significance of the result.
- The **Theil-Sen estimator** calculates the median of the slopes between all pairs of data points, providing an estimate of the trend's magnitude that is less sensitive to outliers than standard linear regression.

We applied the trend analysis to the lemmatized forms of words in the corpora. The parameters for the detection algorithm were set as follows: a minimum total frequency of 50 occurrences within the analyzed period and a significance level of $p < 0.05$. The epoch size for the calculation was set to one month.

3.2 Data Extraction

We conducted the analysis across two languages and two timeframes, resulting in four distinct experimental conditions:

1. **English Trends corpus**, 1-year period (October 2024 — September 2025).
2. **Czech Trends corpus**, 1-year period (October 2024 — September 2025).
3. **English Trends corpus**, 2-year period (March 2023 — February 2025).
4. **Czech Trends corpus**, 2-year period (March 2023 — February 2025).

For each of these four configurations, we extracted the top 100 lemmas with the highest positive trend slope. This produced a total of 400 lemmas for annotation, each accompanied by a sample of concordance lines from the corpus to provide context, along with a relative frequency plot over time.

3.3 Annotation Protocol

Each of the 400 lemmas was classified into one of six categories by both a human annotator and a LLM. The categories were defined to distinguish between genuinely interesting lexical trends and various forms of statistical or data-related noise.

Annotation Categories. The following six labels were used for classification:

Proper Name The word denotes a specific named entity (e.g., a person, location, organization, or product) that is consistently referred to across the provided contexts.

Seasonal Word Usage The word's increased frequency is attributable to a recurring, periodic event (e.g., holidays, seasons, annual sports events) rather than a novel linguistic shift.

Processing Error The trend is an artifact of a data processing issue, such as incorrect lemmatization, tokenization errors, or the word appearing primarily in non-linguistic contexts like captions or boilerplate text.

Source Composition Issue The word's trend is driven by its high frequency in a very limited number of sources or websites, suggesting a topic- or domain-specific surge rather than a broader change in the language.

Interesting Trending Word The word's trend is linguistically or culturally significant, representing a potential neologism, a semantic shift, or a reflection of a noteworthy societal event.

OK The word is genuinely trending and does not fall into any of the error or artifact categories, but its trend is not necessarily of high linguistic interest (e.g., a common word becoming slightly more common).

Human Annotation. A single expert annotator, a linguist with experience in corpus analysis, classified all 400 lemmas. For each lemma, the annotator reviewed concordances from the relevant corpus and time period before assigning one of the six category labels.

LLM Annotation. The same classification task was performed by the GPT-OSS 120B large language model. For each of the 400 lemmas, the model was provided with the same contextual snippets as the human annotator and was prompted to choose one of the six categories. The prompt was structured as follows:

> You are a linguistic researcher. You will inspect the text sample, in which a word deemed to be trending by an automated trend detection system is marked between '<' and '>'. Each of the snippets starts with the source website and the headword. The target word is represented by its lemma, common for all the snippets provided by an external tool, which can make mistakes. The lemma is '[LEMMA]'.
> The relative frequencies within time periods are:
> [RELATIVE FREQUENCIES]
> The snippets are:
> [CONCORDANCE LINES WITH METADATA]
> Respond with one of these answers only: "proper name", "seasonal word usage", "processing error", "source composition issue", "interesting trending word", "ok". The meaning of these answers are:
> - proper name: The target word represents a single specific named entity: a single specific object, such as a person, location, organization, product, etc., that can be denoted with a proper name. It can be abstract or have a physical existence. Only choose this if it is apparent from the provided context that the word denotes the

Table 1: Distribution of Annotation Judgements for Top 100 Trending Words by Language and Time Span. Cohen's κ measures the agreement between the Human and LLM annotators for each dataset.

	Czech				**English**			
	2-year period		1-year period		2-year period		1-year period	
Category	**Human**	**LLM**	**Human**	**LLM**	**Human**	**LLM**	**Human**	**LLM**
OK (Interesting)	**59** (7)	**39** (39)	**57** (6)	**31** (29)	**54** (5)	**23** (18)	**47** (2)	**33** (30)
Proper Name	22	22	10	8	15	17	13	13
Seasonal Word Usage	0	8	24	45	0	2	9	15
Processing Error	18	16	9	8	27	36	22	19
Source Composition Issue	1	15	0	8	4	22	9	20
Cohen's κ	0.436		0.400		0.288		0.602	

same entity in the snippets.
- seasonal word usage: The target word is trending in the period between minm and maxm, but it is trending in an periodical fashion outside of this range.
- processing error: The word in the majority of the text snippets is heavily malformed, or appears mainly outside of normal text, e.g. in a caption or a footer. Note that malformed target lemma is ok, if the word is recognizable within the text snippets, or the tokenization seems to be finer than expected. You do not need to know the word.
- source composition issue: The word is trending in a single website only or the distribution of the websites is very limited, the text is likely originating from a single source. When "processing error" would be also applicable, prefer that.
- interesting trending word: The fact that this word is trending could be linguistically or otherwise interesting.
- ok: The word is trending, but you cannot find any direct fault with it, but it doesn't seem to be particularly interesting.

The model's responses were collected to measure its agreement with the human expert, thereby assessing its capability to perform this nuanced, context-dependent linguistic analysis.

4 Evaluation Results

The results of the manual and automated annotation processes are summarized in Table 1. This section analyzes the distribution of categories assigned by the human annotator, the performance of the LLM, and the level of agreement between them across the four experimental settings.

4.1 Human Annotation Findings

Across all four datasets, the human annotator found that a substantial proportion of the trending words were valid candidates, classified as `OK` or `Interesting`. While Table 1 reports both of these quantities, we count them together in

further analysis due to their subjectivity. This category accounted for 59 % and 57 % of the Czech lemmas and 54 % and 47 % of the English lemmas for the two-year and one-year periods respectively. This demonstrates that the trend detection algorithm is effective at identifying a high number of relevant lexical changes.

The most common sources of noise were `Processing Error` (mostly due to wrong tokenization caused by bad extraction by Justext) and `Proper Name` (which might or might not be desirable to have in the result, depending on the particular use case). Notably, `Seasonal Word Usage` was a major factor in the one-year Czech dataset (24 %) and also significant in the one-year English data (9 %), but was entirely absent in the two-year datasets for both languages. This shows that the longer two-year timeframe smooths out annual periodic spikes, preventing them from appearing as a significant monotonic trend. On the other hand, the shorter one-year period is more susceptible to flagging words that are simply entering their high season. The `Source Composition Issue` was a minor factor for the human annotator, particularly in the Czech data.

4.2 LLM Performance and Annotator Agreement

The agreement between the human annotator and the GPT-OSS 120B model, as measured by Cohen's Kappa, varied significantly across the datasets. The highest agreement was achieved on the one-year English data ($\kappa = 0.602$), indicating substantial agreement. The Czech datasets showed moderate agreement ($\kappa = 0.436$ and $\kappa = 0.400$). The lowest level of agreement was observed for the two-year English data ($\kappa = 0.288$), which is considered only fair.

Several systematic differences in judgment appeared upon closer inspection:

- **Source Composition vs. Genuine Trend:** The LLM consistently overidentified the `Source Composition Issue` category compared to the human annotator. For example, in the two-year English data, the LLM classified 22 lemmas this way, whereas the human identified only 4. The confusion matrices reveal that many lemmas classified as `OK` by the human annotator were labeled as `Source Composition Issue` by the model, suggesting the LLM may be too sensitive to repeated mentions from a few dominant news sources.
- **Seasonal Usage Discrepancy:** A major point of disagreement occurred in the one-year Czech data regarding seasonal words. While the human labeled 24 words as `Seasonal`, the LLM identified 45. The crosstabulation shows that the LLM frequently classified lemmas that the human considered genuinely `OK` (21 instances) as seasonal instead. The model tended overestimate the seasonality effects within the results.
- **Distinguishing Errors:** The poor agreement on the two-year English data was largely driven by the LLM's difficulty in distinguishing genuinely `OK` trends from certain artifacts. Of the 54 lemmas the human marked as `OK`, the LLM agreed on only 19, classifying 16 as `Processing Error` and 14 as `Source Composition Issue`.

Despite these issues, both annotators showed relatively strong agreement in identifying `Proper Name` across all datasets, suggesting this is a more straightforward category for both human and LLM to identify from context. The model also successfully identified a similar number of processing errors in the English one-year dataset, contributing to the high kappa score in that setting.

5 Conclusion

We provided a quantitative evaluation of an automated trend detection system available within Sketch Engine, quantifying the nature of the results and the noise present within it. The analysis of the top 100 trending lemmas in Czech and English across one- and two-year periods confirms that the trend detection feature in Sketch Engine is an effective tool for identifying candidate words for lexicographical review, with a human annotator validating approximately half of the suggestions as genuinely trending words. Interesting finding is the role of the analysis timeframe: a longer two-year period successfully removed the periodic noise of seasonal words, which represents a significant part of the result list in the shorter one-year analysis.

Our evaluation of the GPT-OSS 120B model against a human expert quantifies the current state of LLMs for linguistic filtering. While the model achieved significant agreement in one condition ($\kappa = 0.602$), its performance was not consistent across the board. We identified some systematic biases, particularly the model's tendency to over-classify lemmas as `Source Composition Issue` and `Seasonal Word Usage`, often mislabeling what a human expert considered a valid trend. This suggests that while LLMs can replicate human judgment on more clear-cut categories like `Proper Name`, they struggle with the distinction between a concentrated, topic-specific surge and a broader linguistic shift, or perhaps this is only a weakness of the particular model or the specifics of the prompting strategy we used. To improve the result confidence, multiple human annotations would strenghten our analysis and allow us to distinguish between noise and true biases within the model, or possible uncertainty within our annotation manual.

In practice, this shows that while automated trend detection provides a useful starting point, unfiltered lists are noisy, and the choice of time span is a critical parameter. While LLMs show promise, they are not yet a direct replacement for expert human judgment on this task. Future work could explore more sophisticated prompting or fine-tuning to reduce these biases, potentially enabling a hybrid workflow where an LLM serves as a first-pass filter to reduce the manual workload for lexicographers and linguists.

Acknowledgments. This work has been partly supported by the Ministry of Education, Youth and Sports of the Czech Republic within the LINDAT-CLARIAH-CZ project LM2023062.

References

1. Herman, O., Jakubíček, M., Kraus, J., Suchomel, V.: 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 (2025)
2. Kilgarriff, A., Baisa, V., Busta, J., Jakubicek, M., Kovar, V., Michelfeit, J., Rychly, P., Suchomel, V.: The sketch engine: ten years on. Lexicography **1**(1), 7–36 (2014)
3. Michelfeit, J., Pomikálek, J., Suchomel, V.: Text tokenisation using unitok. In: Horák, A., Rychlý, P. (eds.) RASLAN 2014. pp. 71–75. Tribun EU, Brno, Czech Republic (2014)
4. Pomikálek, J.: Removing boilerplate and duplicate content from web corpora. Ph.D. thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic (2011)
5. Rychly, P.: Manatee/bonito–a modular corpus manager. RASLAN 2007 Recent Advances in Slavonic Natural Language Processing p. 65 (2007)
6. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: New methods in language processing. pp. 154–164. Routledge (2013)
7. Šmerk, P.: Fast morphological analysis of czech. RASLAN 2009 Recent Advances in Slavonic Natural Language Processing p. 13 (2009)
8. Trampus, M., Novak, B.: Internals of an aggregated web news feed. In: 15th Multiconference on Information Society. pp. 221–224 (2013)