From Word of the Year to Word of the Week:

Daily-updated Monitor Corpora for 25 Languages

Ondřej Herman[1,2], Miloš Jakubíček[1,2], Jan Kraus[1], Vít Suchomel[1,2]

[1]Lexical Computing, Brno, Czech Republic
[2]Natural Language Processing Centre, Masaryk University, Brno, Czech Republic
firstname.lastname@sketchengine.eu

Abstract

This paper introduces a long-term, privately funded programme to collect time-stamped monitor Trend Corpora in a wide range of languages, designed to study linguistic trends and language change over time. Accessible via the Sketch Engine platform, the corpora range in size from 3 million tokens (Irish) to 100 billion (English). 25 languages are covered – including Arabic, English, French, German, Italian, Polish, Portuguese, and Spanish – with ten more to be added soon.

Corpus texts come from global websites providing RSS/Atom feeds, mostly news, covering content from as early as 2014. New articles – up to 180,000 on weekdays – are collected daily and updates are published twice a week. Processing includes text cleaning, de-duplication, and linguistic annotation. The project builds on the JSI Newsfeed Corpus (Krek et al., 2017), but since 2021 for English and 2023 for other languages it has expanded independently in scope and data sources.

Trend corpora in Sketch Engine support diachronic analysis across multiple time frames and integrate with features like concordance search and Word Sketch. The paper also presents feed activity statistics and showcases examples of functionality offered by Trend Corpora such as neologism detection, word sense shift analysis, and timeline-based analysis of trending words and phrases.

Keywords: monitor corpus; web corpus; trend analysis; neologism detection; word sense shift analysis

1. Introducing Trend Corpora

To examine linguistic trends, neologisms, and language change over time, it is essential to provide a corpus with high-quality, timestamped texts and a user interface to browse the corpus. Important features of high quality data include extensive temporal coverage, reliably determined timestamps, and thorough removal of HTML boilerplate and other unwanted content in the case of data obtained from the web.

Concerning the usability of the corpus interface, it should offer timestamp-based analyses of language change, function efficiently with both small and very large corpora and provide an intuitive, user-friendly, responsive interface. This paper addresses both aspects of Trend Corpora: the data itself and the user interface through which it can be queried.

Trend Corpora discussed in this article are published and regularly updated in Sketch Engine.[1] They are accessible through a graphical user interface and via an API.[2]

By automating the procedures for collecting texts and corpus building, which is described in Section 3, we were able to provide fresh data to all supported languages twice a week now. Applications of the existing Sketch Engine interface, with a focus on its use by lexicographers and linguists, are illustrated in Section 4. The planned future developments of Trend Corpora are discussed in concluding Section 5.

2. Related Work

This section provides an overview of different existing projects with comparable objectives.

2.1 News on the Web Corpus

The NOW (News on the Web) corpus (Davies, 2017)[3] is a large-scale English monitor corpus, currently comprising approximately 22.3 billion words from online newspapers and magazines, covering the period from 2010 to the present. It is updated daily, adding 6–10 million words each night, which amounts to 200–220 million words monthly from roughly 300,000–400,000 articles, or about 1.8–2.0 billion words annually. Data is gathered from approximately 1,200 news websites, with new articles processed automatically each day, cleaned using JusText, lemmatized and tagged, and integrated into a relational database. This pipeline is reported to do the processing in 4–5 hours.

The corpus is designed to capture recent linguistic developments in English, enabling exploration of new or emerging terms (e.g. Brexit, pandemic, fake news) that may not be present in older corpora. It allows frequency analyses by month or year, pattern-based queries (e.g. data NOUN, -gate, -phobia), and comparative analyses across countries or time periods. Users can also create virtual corpora filtered by topic, source, or date – for example, constructing a corpus of articles from September 2015 discussing refugees in Europe, and extracting keyword lists from such custom datasets.

2.2 Corpus of Contemporary American English

The Corpus of Contemporary American English (COCA) (Davies, 2010) is a 1-billion-word corpus spanning 1990–2019 with consistent annual genre balance. It includes diverse registers – spoken, fiction, magazines, newspapers, academic texts, blogs, and subtitles – making it suited for diachronic and genre-based linguistic research. COCA enables detailed frequency, collocation, and syntactic pattern analysis across time and context.

2.3 JSI Newsfeed Corpus

A related development is based on the JSI Newsfeed (Trampus & Novak, 2013), crawled by the Jožef Stefan Institute, providing a real-time stream of multilingual semantically-enriched news articles obtained from RSS feeds. The JSI Newsfeed was adapted to the JSI

[1] https://www.sketchengine.eu/
[2] https://www.sketchengine.eu/apidoc/
[3] https://www.english-corpora.org/now/

Newsfeed Corpus (Krek et al., 2017) for use within Sketch Engine. The corpus is available in Arabic, Catalan, Czech, German, English, Finnish, French, Croatian, Hungarian, Italian, Korean, Dutch, Polish, Russian, Spanish, Serbian, Swedish. The project provided monitor corpora for 13 languages, updated daily. However, JSI Newsfeed finally ceased to provide data in 2022, and no further updates will be made.

The JSI Newsfeed corpora are preserved in Sketch Engine. The Trend corpora presented in this article aim to extend the data and provide a comparable resource spanning a longer time period. Many of the Trend corpora already contain the JSI Newsfeed data – retaining also the most important metadata such as the source URL, publication date, and geographical information – to enable diachronic research over a much longer timespan. We are currently working on integrating the remaining language variants.

2.4 Google Books Ngrams

One of the largest resources for studying language change is the Google Books Ngram Corpus (Lin et al., 2012), which, in its English variant, comprises 1–5-grams extracted from over 8 million books or about 6 % of all books ever published between 1500s and 2022. N-gram frequencies are available for download up to the year 2008; subsequent data can be accessed via the web interface.

The n-grams are provided with temporal annotation with a granularity of one year. Key limitations include the narrow context, the absence of metadata and access to the underlying texts, and the uncertainty surrounding future updates – the most recent data in the corpus dates to 2022. This corpus is also available in French, Spanish, German, Chinese, Russian, and Hebrew.

2.5 Other Monitor Corpora

Diachronic corpora providing cutting-edge data tend to be short-lived or focused on specific languages, and typically limited in scope. Examples include the Trendi corpus of Slovene (Kosem, 2022), the Irish National Monitor Corpus (Ó Meachair et al., 2022), or the now-defunct service provided by www.monitorcorpus.com.

3. Data

In this chapter, we describe the identification of possible data sources via news feeds, the procedure of obtaining and processing texts from web pages published through the feeds, and compiling the result corpora using Sketch Engine.

3.1 Feed Exploration

Feed-based corpora are built using RSS (Really Simple Syndication) feeds, which work in the background to collect content from websites. For this process to work, a website must provide an RSS file – usually with a `.rss` or `.xml` extension. Since offering an RSS feed is optional, not all websites include one, and those that do not are left out of the corpus.

Whenever a website publishes new content, that update is reflected in its RSS feed. These feeds are regularly checked for updates, and once new content appears, it can be downloaded along with metadata such as the publication date, title, and description.

Most of the websites used in this approach are news sites, as they tend to publish the largest volume of content.

In most cases, RSS feeds are found manually. This is mainly because the availability of feeds depends on the language and how well it is supported online. For widely used languages like English, there are often complete lists of RSS feeds that can be downloaded and processed using simple scripts. However, such lists are usually unavailable for lower-resourced languages, which makes it necessary to search for feeds by hand.

The process typically starts with a set of topics that are common in TenTen Corpora[4]. A researcher uses these topics to search for websites using one or more search engines–it's helpful to try different ones, as they may return different results.

Each website is then manually checked to see if it provides an RSS feed. Sometimes, there is a standard RSS icon or a link labeled *RSS*; other times, the feed might be hidden in the page source or tucked away in a less obvious section of the site.

At this moment, the topic, genre, and location related to the articles published by the feed are determined by the person checking the website. This information is later added as metadata of texts coming from the feed, provided that there is a single topic, genre, or location common to most articles there. The exact methodology was described in Suchomel et al. (2022).

This process is currently manual to allow for better quality control. Automated methods might pull in all feeds from a site, even if some are not relevant or needed. Doing it manually helps ensure that only useful and appropriate feeds are included.

3.2 Obtaining Texts from Feeds

The crawler runs as a batch job every four hours in two stages. In the first stage, each web feed is visited and all previously unseen pages are added to the download schedule.

During the second stage, the crawler attempts to download the scheduled web pages. The pages are downloaded concurrently from up to 500 servers. However, requests to any single server are made sequentially and never more frequently than once every five seconds, to avoid overloading it. The crawler waits up to 120 seconds for a server to respond with the web page content. If the server fails to respond within the allotted time, or returns an error, the web page is rescheduled for the crawler's next run. After three failed attempts, the page is removed from the schedule, and no further download attempts are made in any of the following runs. After one hour, any remaining downloads are terminated, and the crawler shuts down.

Diachronic annotations for the downloaded web pages are based on the *published* timestamp provided by the web feed. If the *published* timestamp is unavailable, the time at which the

[4] https://www.sketchengine.eu/blog/topics-and-genres-in-corpora/

page was first observed in the feed is used instead. As a validation step, clearly incorrect timestamps [5] are replaced with the download time.

3.3 Text Processing

Text extraction, cleaning, segmentation, annotation, and indexing follow the process described by Kilgarriff et al. (2014): Paragraphs are extracted from original web pages using JusText (Pomikálek, 2011). The tool also filters out paragraphs that are too short (e.g. page navigation and other HTML boilerplate) or chunks not containing words in the target language based on a built-in heuristic. Duplicate and near-duplicate paragraphs are removed using Onion (Pomikálek, 2011).

3.4 Indexing in Sketch Engine

The final corpora are compiled and indexed in Sketch Engine (Kilgarriff et al., 2014). Since the compilation of the largest 100-billion word corpus requires three days, the update frequency for all corpora has been set to twice a week.

The entire worklow – including source downloading, text processing, corpus compilation, deployment, and data backup – is driven using Makefiles (invoked by `make`), ensuring that all subsequent procedures can be rerun efficiently when a tool is changed (e.g. a POS tagger model) or when new source data is downloaded. The entire process is fully automated and executed as a regular scheduled task using `cron`.

3.5 Resulting Trend Corpora

The properties of the currently published Trend Corpora are summarized in Table 1. The total size of each corpus, including texts from the JSI Newsfeed corpora collected between 2014 and 2022 by Krek et al., is also shown. The last two columns indicate the availability of additional, semi-manually identified text types.

3.6 Statistics on Feed Activity

Table 2 presents the number of tokens and documents added between July 1, 2024, and June 30, 2025, along with the number of websites from which the pages were obtained, along with the number of source websites for each language variant of the corpus.

A larger set of RSS feeds has been collected for widely spoken languages, due to their stronger online presence. These feeds also tend to yield more content. As a result, they produce greater quantities of both documents and tokens.

On the other hand, less widely spoken languages such as Irish or Maltese have a more limited online presence. Consequently, fewer RSS feeds are available, and the volume of collected text is significantly smaller. These size differences stem primarily from content availability rather than technical limitations.

[5] Such as malformed dates, or those set in the future or more than two years in the past at the time of download.

Table 1: Overview of languages of Trend Corpora as of July 2025 – languages marked with an asterisk denote Trend Corpora merged with JSI corpora

Language	**start of collection**	**collected until July 2025**		**total size incl. JSI texts collected in 2014–2022**		**topic, genre**	**location**
		documents	**tokens**	**documents**	**tokens**		
Arabic*	Feb '23	5,167,129	1,365,325,316	33,965,431	7,699,143,384		
Catalan	Mar '23	153,666	81,431,128	153,666	81,431,128	✓	
Czech*	Feb '23	2,203,721	1,083,277,302	7,014,200	2,705,490,161	✓	✓
Danish	May '23	483,639	153,426,710	483,639	153,426,710	✓	
Dutch	May '23	1,074,106	408,467,106	1,074,106	408,467,106	✓	
English*	Jun '21	24,793,691	14,049,099,262	260,075,574	99,668,563,911	✓	✓
Estonian	Feb '20	1,106,965	286,713,575	1,106,965	286,713,575	✓	
French	May '22	2,042,464	1,158,086,284	2,042,464	1,158,086,284		
German	Dec '22	6,658,163	2,370,156,271	6,658,163	2,370,156,271	✓	
Greek	Oct '23	3,921,354	1,016,170,981	3,921,354	1,016,170,981	✓	
Hebrew	Aug '23	684,037	334,381,122	684,037	334,381,122	✓	
Hungarian	Oct '23	1,675,497	562,542,800	1,675,497	562,542,800	✓	
Italian*	Jan '23	6,746,747	2,857,538,871	32,124,715	11,270,141,041	✓	
Irish	Aug '23	4,855	3,139,030	4,855	3,139,030	✓	
Maltese	Aug '23	39,810	11,342,848	39,810	11,342,848		
Norwegian Bokmål	Sep '23	305,087	113,358,245	305,087	113,358,245	✓	
Norwegian Nynorsk	Sep '23	25,490	12,423,910	25,490	12,423,910	✓	
Persian	Aug '23	1,531,630	541,814,380	1,531,630	541,814,380	✓	
Polish	Feb '23	2,538,742	1,095,201,030	2,538,742	1,095,201,030		
Portuguese	Apr '23	2,887,738	1,210,087,551	2,887,738	1,210,087,551	✓	
Russian	Apr '23	10,273,716	2,694,195,098	10,273,716	2,694,195,098	✓	
Slovak	Aug '23	1,085,561	356,104,312	1,085,561	356,104,312	✓	
Slovene	Aug '23	591,936	206,904,510	591,936	206,904,510	✓	
Spanish	Dec '22	4,271,491	2,201,262,585	4,271,491	2,201,262,585		
Ukrainian	May '22	4,496,548	1,174,970,729	4,496,548	1,174,970,729		
Total: 25		**84,763,783**	**35,347,420,956**	**379,032,415**	**137,335,518,702**	**19**	**2**

3.7 The Most Represented Web Domains per Language

Table 3 presents the most prominently represented web domains for each language-specific corpus. The first column specifies the language of the corpus. The *Webs* column indicates the total number of distinct web domains included in the corpus. For each corpus, four largest contributing web domains are listed in the *Web Domain* column. The *Proportion (%)* column reports the relative contribution of each domain to the overall corpus size.

To ensure a corpus sufficiently represents multiple text types (e.g. genres and topics), it is essential to source its content from diverse origins. However, as shown in the table, this is not the case for Trend Corpora in Irish, Maltese, and Nynorsk. The primary reason is the relative scarcity of web feeds available in these languages compared to others.

Corpus	Added Tokens	Added Documents	Active Feeds
Arabic	612,723,407	2,059,700	260
Catalan	66,587,955	114,949	49
Czech	490,590,699	912,083	690
Danish	68,094,443	199,105	119
Dutch	157,028,190	388,009	137
English	4,645,146,400	7,768,628	8,047
Estonian	57,966,679	195,457	263
French	406,283,300	663,897	206
German	1,094,981,747	2,787,513	428
Greek	616,881,978	2,143,753	143
Hebrew	188,548,557	345,784	115
Hungarian	333,504,488	868,681	142
Irish	1,844,550	2,629	8
Italian	1,275,326,540	2,733,267	839
Maltese	5,568,611	17,343	7
Norwegian Bokmål	65,409,547	162,204	82
Norwegian Nynorsk	7,181,463	13,774	72
Persian	323,364,752	856,460	25
Polish	481,303,215	1,009,522	310
Portuguese	572,402,169	1,246,785	501
Russian	1,287,876,786	4,533,403	622
Slovak	197,289,002	545,316	203
Slovene	114,448,592	294,798	113
Spanish	976,364,646	1,669,675	216
Ukrainian	468,423,257	1,475,235	261

Table 2: Tokens, documents and active Web feeds over the past year (July 1, 2024–June 30, 2025)

3.8 Data Addition and Removal

New items can be added to the list of subscribed feeds as needed. This has previously been done for the English, German, Czech, and Italian corpora, as additional feeds were discovered after data collection had begun. It is acknowledged that adding highly productive feeds – especially those covering genres or topics previously underrepresented in the corpus – may influence trend analyses. Users should always examine a trend candidate within its corpus context and consult relevant metadata – such as the frequency distribution of source feeds associated with concordance hits – to verify whether the candidate represents a genuine trending word. This is supported by functionality available in Sketch Engine.

Language	Webs	Web Domain	%
Arabic	3,855	alaraby.co.uk	5.39
		aawsat.com	5.19
		aljazeera.net	2.86
Catalan	67	ara.cat	15.79
		arabalears.cat	13.87
		xcatalunya.cat	8.95
Czech	7,474	medium.seznam.cz	5.12
		novinky.cz	2.78
		seznamzpravy.cz	2.73
Danish	174	ekstrabladet.dk	10.94
		via.ritzau.dk	10.14
		politiken.dk	5.98
Dutch	154	nrc.nl	9.13
		tweakers.net	5.07
		voetbalnieuws.be	3.29
English	319,269	dailymail.co.uk	2.75
		theguardian.com	1.89
		independent.co.uk	1.37
Estonian	567	pmo.ee	32.96
		err.ee	8.93
		ohtuleht.ee	6.46
French	400	lemonde.fr	5.48
		la-croix.com	3.85
		ledevoir.com	3.36
German	621	tag24.de	2.39
		krone.at	1.94
		noen.at	1.83
Greek	164	in.gr	3.65
		gazzetta.gr	3.18
		protothema.gr	2.96
Hebrew	172	israelhayom.co.il	15.17
		maariv.co.il	8.98
		one.co.il	6.84
Hungarian	181	index.hu	5.58
		origo.hu	4.49
		mandiner.hu	4.13
Irish	10	tuairisc.ie	88.08
		meoneile.ie	5.39
		paroistebailemhuirne.ie	3.26

Language	Webs	Web Domain	%
Italian	12,216	fanpage.it	1.89
		ilrestodelcarlino.it	1.57
		unionesarda.it	1.49
Maltese	10	newsbook.com.mt	48.94
		one.com.mt	30.74
		netnews.com.mt	13.38
Norwegian	91	dagsavisen.no	13.51
Bokmål		nrk.no	12.93
		aftenposten.no	10.16
Norwegian	77	nrk.no	36.61
Nynorsk		dagogtid.no	17.79
		vl.no	9.06
Persian	30	fararu.com	9.33
		tn.ai	8.04
		mehrnews.com	7.85
Polish	539	sport.pl	2.45
		rp.pl	2.14
		sportowefakty.wp.pl	1.89
Portuguese	950	noticiasaominuto.com	5.72
		cnnbrasil.com.br	4.23
		uol.com.br	3.26
Russian	805	rg.ru	2.58
		ura.news	1.83
		lenta.ru	1.78
Slovak	855	sme.sk	4.23
		hnonline.sk	3.28
		aktuality.sk	3.19
Slovene	117	rtvslo.si	12.12
		24ur.com	9.23
		siol.net	8.75
Spanish	314	lavoz.com.ar	3.84
		okdiario.com	3.80
		mundodeportivo.com	3.49
		elcomercio.pe	3.15
Ukrainian	552	24tv.ua	6.29
		uk.wikipedia.org	5.50
		unian.ua	4.30

Table 3: Most represented web domains per language

In addition to removing inactive sources, feeds or documents from specific websites are excluded in two cases: (1) if the content is machine-generated or malformed,[6] or (2) if a copyright holder requests removal–though this has not occurred to date.

4. Trend Corpora in Sketch Engine

This chapter shows six use cases of Trend Corpora facilitated by functions of Sketch Engine. The usefulness for lexicographers or general linguists is explained.

4.1 Neologism Detection

Sketch Engine provides a Trends[7] feature for selected corpora, which facilitates the identification of words exhibiting significant changes in frequency over time, thereby supporting diachronic linguistic analysis. This functionality can aid in the detection of neologisms – newly emerging words in a language.

The tool is particularly valuable for lexicologists, historians, and other researchers seeking to ensure that language resources remain contemporary and reflective of actual usage. Moreover, identifying neologisms has practical applications in areas such as machine translation, speech recognition, and sentiment analysis, where up-to-date lexical data enhances performance.

From a broader perspective, tracking neologisms offers insights into societal change and evolving cultural values, which may inform marketing strategies and communication practices within commercial sectors.

As illustrated in Figure 1, the Trends result page presents a table comprising four primary columns: Lemma, Trend, Frequency, and Sample, accompanied by additional action options for each row (hidden under the three dots icon), specifically Concordance and Frequency. The Lemma column lists the identified lemmas, while the Trend column displays a numerical trend value along with an arrow indicating a significant increase or decrease in usage. The Frequency column provides the absolute frequency of each lemma, and the Sample column offers a graphical representation of the lemma's frequency variation over time, typically covering the most recent years.

For the purpose of neologism detection, the researcher manually reviews the Trends results and, drawing on their linguistic expertise, identifies potential neologism candidates. This initial stage of the process is typically the most time-intensive, often requiring several hours to complete. It may be necessary to experiment with various parameter settings to refine the results; for example, in our analysis, many emergent words exhibited relatively low absolute frequencies–typically in the range of a few hundred occurrences.

To evaluate unfamiliar items, the Concordance function can be used to inspect contextual sentence examples. This helps determine whether a given word represents a plausible neologism or merely constitutes an outlier with little semantic coherence in context. If the word appears meaningful and contextually valid, the researcher may proceed to search for its occurrence in external sources using web search engines.

[6] e.g., due to broken character encoding
[7] https://www.sketchengine.eu/guide/trends/

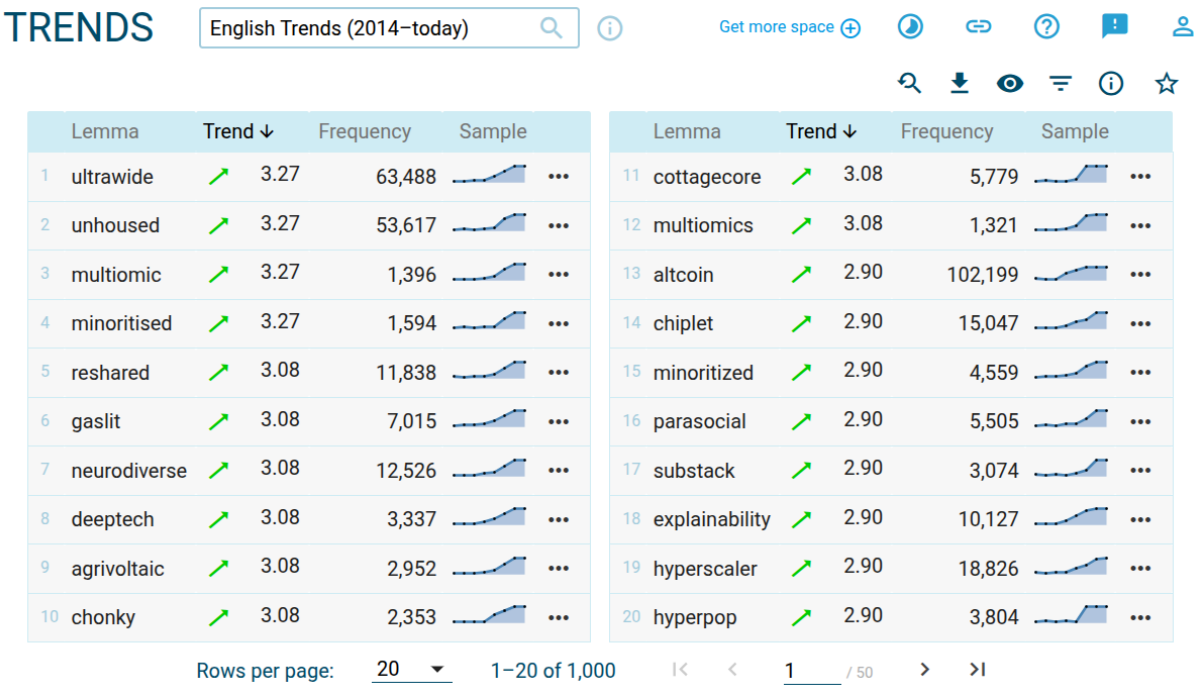

Figure 1: English Trends corpus – results for minimum frequency of 1000, p-value 0.05, and only positive trend enabled (calculated on July 8, 2025)

Quarter	Value	Quarter	Value	Quarter	Value
2018 Q1	1	2018 Q2	1	2018 Q4	1
2019 Q1	3	2019 Q2	1	2019 Q4	11
2020 Q1	12	2020 Q2	17	2020 Q3	75
2020 Q4	210	2021 Q1	124	2021 Q2	67
2021 Q3	246	2021 Q4	289	2022 Q1	594

Table 4: Sample quarterly counts from 2018 Q1 to 2022 Q1 for the word *hyperpop*

As shown in Table 4, the frequency of the word *hyperpop* – a kind of pop music – began to increase significantly starting in 2019. According to the relevant Wikipedia article,[8] the term gained popularity in that year, particularly on the social media platform TikTok. This observation aligns with the trends identified in our corpus data. The Frequency function in Sketch Engine enables the analysis of a word's development over time in terms of absolute frequency. Furthermore, the frequency data can be exported from Sketch Engine in various formats to support additional analysis. The word *hyperpop* is currently not included in any major dictionaries, such as the Oxford English Dictionary, Merriam-Webster, or Dictionary.com. Its absence suggests that, despite several years of usage, it remains a candidate for potential inclusion as a neologism in future dictionary updates.

[8] https://en.wikipedia.org/wiki/Hyperpop

4.2 Timeline-Based Analysis of Trending Words and Phrases

The Concordance function in Sketch Engine includes a Timeline feature[9] that visualizes the frequency of a word or phrase over time. The function needs to be enabled on the Concordance results page (as shown in Figure 2). This tool enables linguists to investigate patterns of language change, including the emergence of new words, semantic shifts, and the decline or obsolescence of older terms. Such temporal analyses can reveal how linguistic trends are shaped by social, political, or technological developments. Notably, sudden increases in frequency may correspond with specific real-world events, providing valuable insights for socio-linguistic research.

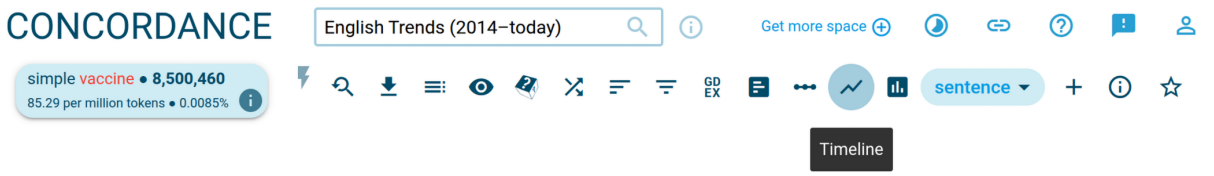

Figure 2: Sketch Engine Concordance – Timeline function button

Figure 3 presents the Timeline for the word *vaccine*, illustrating a marked increase in frequency beginning in 2020, which corresponds to the onset of the global COVID-19 pandemic. This trend reflects the widespread use of the term in public discourse, particularly in news media. In the graph, the blue line indicates the relative frequency per million words, while the grey bars represent the absolute frequency. The entire visualization can be downloaded by users in either `.png` or `.svg` format, allowing for further editing and integration into research outputs.

Another illustrative example – as shown in Figure 4 – is the phrase *presidential elections*, which clearly aligns with real-world events. In the United States, presidential elections were held in 2016, 2020 and 2024. The Timeline graph shows distinct spikes in frequency during the month of November in all the years. This pattern demonstrates the sensitivity of corpus data to major political events and highlights how linguistic frequency reflects societal focus. Such insights are useful not only for linguistic analysis but also for researchers in media studies, political communication, and discourse analysis, as they offer empirical evidence of when and how certain topics gain prominence in public discourse. Moreover, detecting such regular temporal patterns can support automated event detection, temporal segmentation in large corpora, or even training data selection for domain-specific natural language processing tasks.

The final example in Figure 5 displays the Timeline for the Italian phrase *crisi di governo* (government crisis) in the Italian Trends corpus, which shows clear spikes in frequency in 2019, 2021 and 2022 – years marked by political instability in Italy. For instance, in 2021, Matteo Renzi's Italia Viva party withdrew support from Giuseppe Conte's government, prompting Conte's resignation and the formation of a new government under Mario Draghi. In 2022, the Five Star Movement withdrew support from Draghi's administration, resulting in his resignation and the dissolution of parliament. These frequency peaks effectively mirror the timing of real political events, demonstrating the corpus's responsiveness to national developments.

[9] https://www.sketchengine.eu/timeline-language-use-over-time/

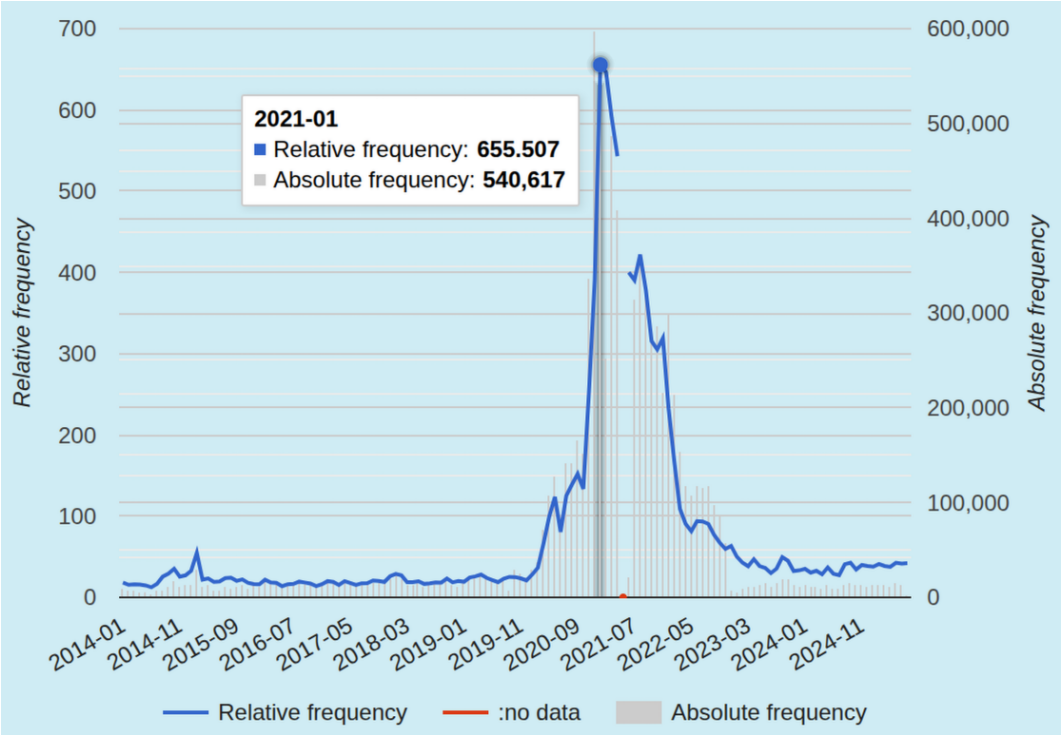

Figure 3: Timeline for the word *vaccine*. The marked decrease in its absolute frequency observed since 2023 can be attributed to the consolidation of corpus data sources, whereby the JSI Newsfeed was replaced with the new sources incorporated into corpus Czech Trends after November 2022.

4.3 Word Sense Shift Analysis

In addition to identifying trending words, the corpora can be used to study subtler changes in word usage and to identify trending senses.

This functionality is particularly valuable for lexicographic work. Compiling a dictionary is a complex and labor-intensive task. Since language is a dynamic and evolving system, a dictionary can never be considered truly complete. While identifying new headwords is already challenging, updating existing word senses of established entries often demands a comparable level of effort, as it requires examining substantial corpus evidence to reliably confirm the emergence of a new sense.

To facilitate the identification of novel word senses, Sketch Engine's trend detection can be combined with word sense clustering to reveal senses that are increasing in prominence over a specific time period using the method described in Herman (2025).

For example, the English noun *rag* is now strongly associated with artificial intelligence, as illustrated in Figure 6, compared to its past usage. The stacked plots represent the relative frequency of specific word senses over time. Current usage is dominated by the initialism *rag* (retrieval-augmented generation), referring to a technique used for querying large language models, whereas the traditional sense – referring to a piece of cloth – has become relatively uncommon.

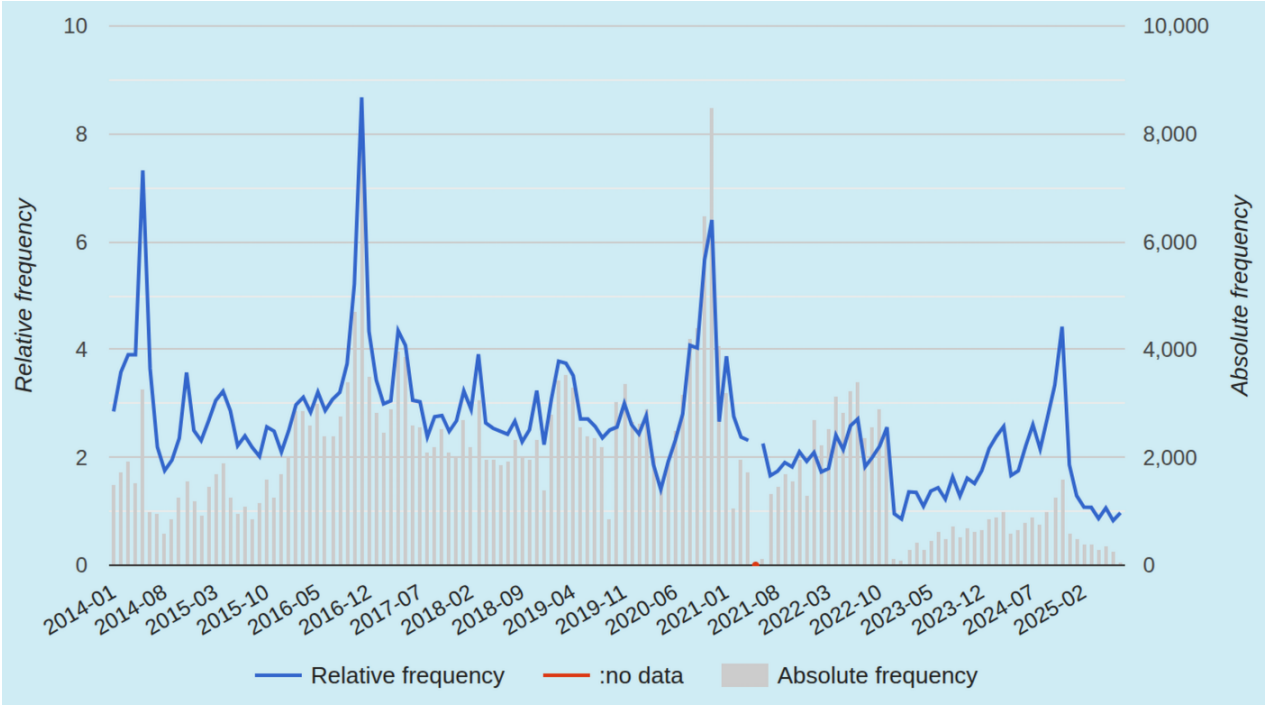

Figure 4: Timeline for the phrase *presidential elections*

The Czech noun *motorista* provides another illustrative example. The conventional meaning of the word is the equivalent of English words *a motorist, driver*. Currently, as can be seen in Figure 7, the majority of usages refer to a member of a far-right political party.

4.4 Wordlist Discovery

In the Wordlist feature, timelines[10] are visualized as small graphs next to each word in the result display. These graphs illustrate frequency trends over time and can be generated using any criteria applied to the source wordlist. When working with large, multibillion-word corpora, graph generation may require a longer processing time, often up to several dozen seconds.

Figure 8 presents a sample wordlist from the English Trends corpus, featuring the words woman, president, and health. The timeline graphs for each word are displayed collectively on a single page, facilitating quick and effective comparison. Notably, the word *health* – similar to *vaccine* – shows a marked increase in frequency in early 2020, corresponding to the onset of the COVID-19 pandemic.

Displaying multiple timelines on a single page offers several advantages for linguistic research. It allows for efficient visual comparison and immediate recognition of patterns, such as parallel frequency spikes or shifts in usage. This setup supports faster exploratory analysis, enabling researchers to refine hypotheses with greater ease. Additionally, it enhances the clarity and accessibility of findings in both pedagogical and presentational contexts.

[10] https://www.sketchengine.eu/timeline-language-use-over-time/

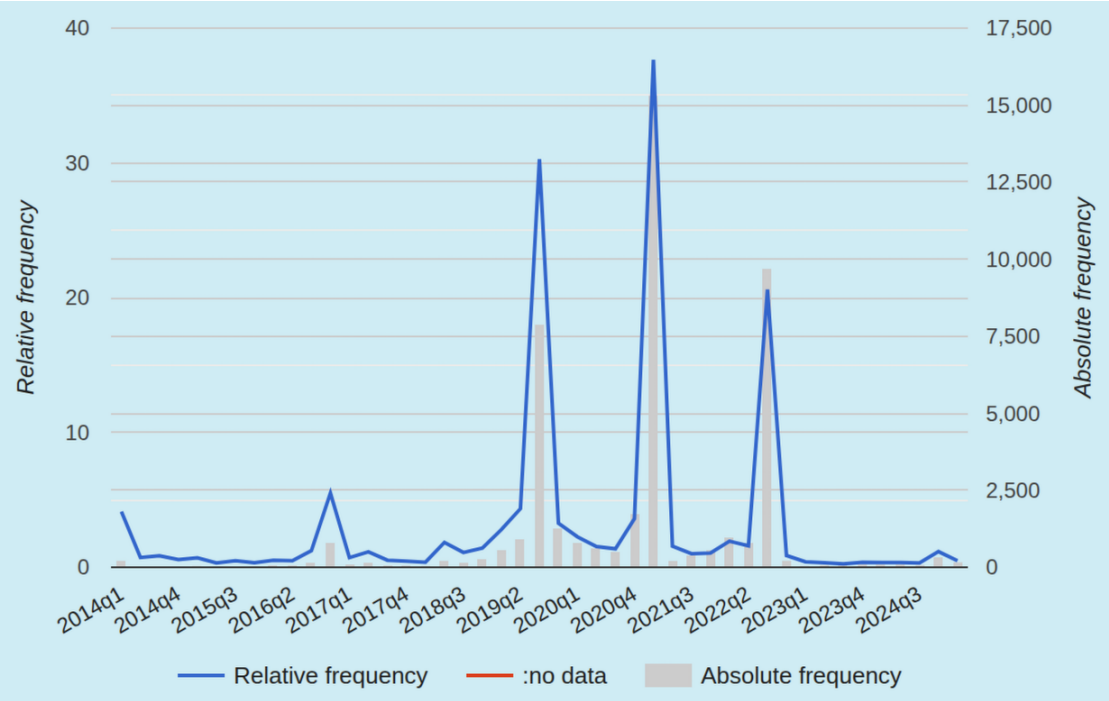

Figure 5: Timeline for the Italian phrase *crisi di governo* (government crisis)

4.5 Collocational Behaviour of Words – Word Sketch

The Word Sketch function[11] analyzes a word's collocates and surrounding context, providing a concise, one-page summary of its grammatical and collocational behavior. The output is structured into categories known as grammatical relations, which include patterns such as words functioning as subjects or objects of a verb, or as modifiers. The inclusion of specific collocations in the analysis is determined by predefined rules encoded in the sketch grammar.

Each collocation can be further explored using other functions – Concordance, Word Sketch, or Thesaurus – via the menu accessed by clicking the three-dot icon beside each collocate, as shown in Figure 9. As illustrated in Figure 10, the collocation *government shutdown* was examined in the Concordance view, enabling the researcher to conduct a more detailed analysis, such as observing its frequency pattern over time using the Timeline feature.

4.6 Word Usage Change in Time

Sketch Engine provides the Word Sketch Difference[12] function for comparing collocational behavior. One useful application is the comparison of subcorpora from different time periods. By leveraging metadata, users can create subcorpora for specific years and analyze changes in collocation patterns. This is particularly valuable for lexicologists, as it helps identify shifts in usage or the emergence of new collocates, offering insights into lexical development and semantic change.

[11] https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/
[12] https://www.sketchengine.eu/guide/word-sketch-difference-compare-words/

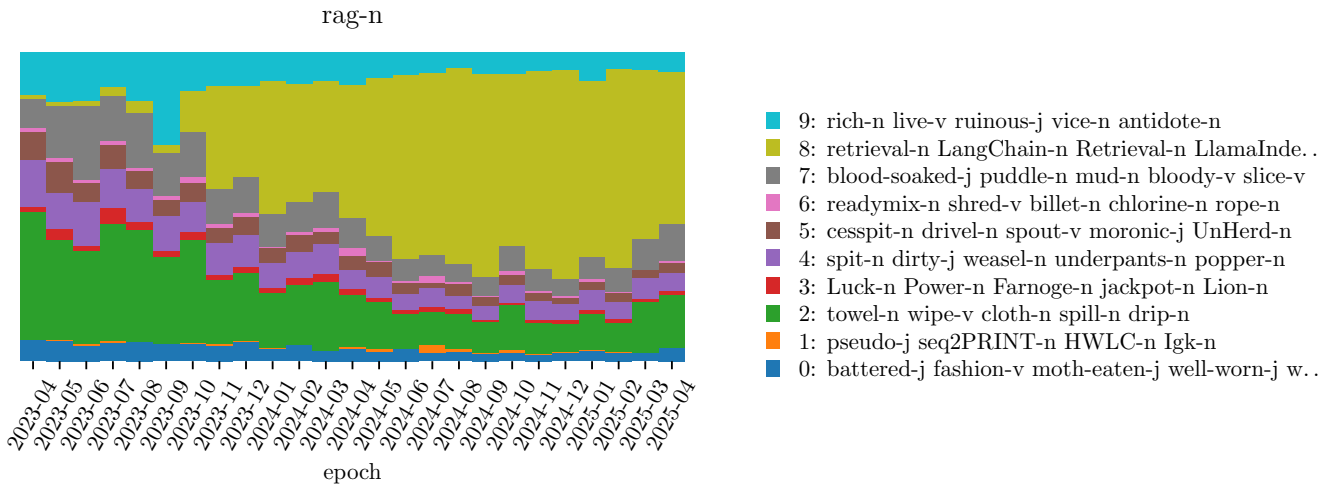

Figure 6: Diachronic word sense distribution of the English noun *rag* – Current usage is dominated by the initialism *rag* (retrieval-augmented generation), represented by trending sense 8 (in pear colour) referring to a technique used for querying large language models, whereas the traditional sense – referring to a piece of cloth – has become relatively uncommon.

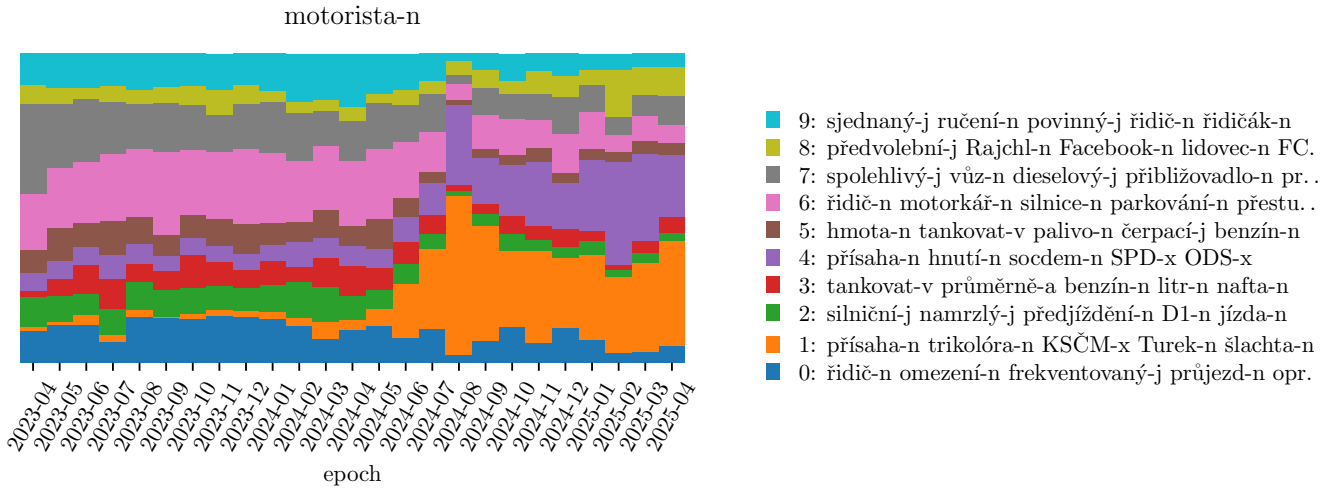

Figure 7: Diachronic word sense distribution of the Czech noun *motorista* (a motorist, driver) – Note that the currently trending senses 1 and 4 (represented in orange and violet) denote a member of a Czech far-right political party that secured a seat in the European Parliament in 2024 and has remained in the public spotlight since then.

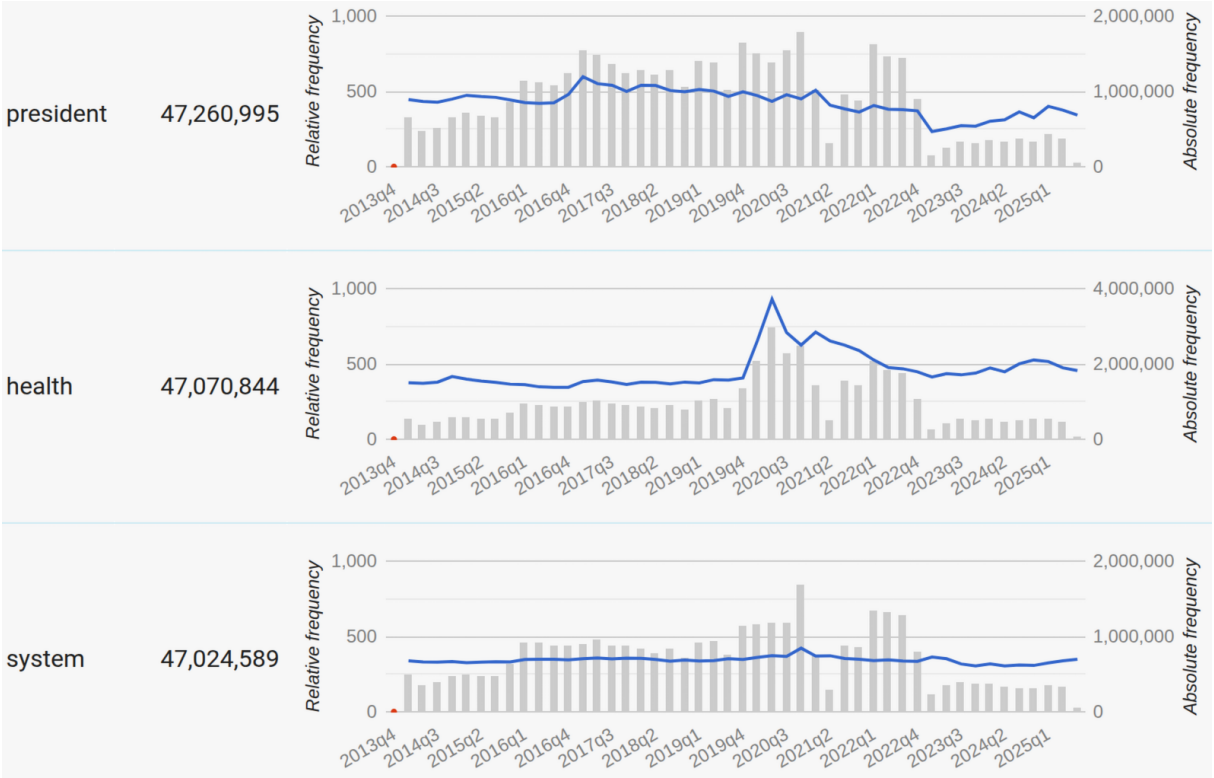

Figure 8: Sketch Engine Wordlist – sample cut-out with three words showing built-in Timelines

Figure 11 displays the Word Sketch Difference for the word *llama*, based on subcorpora from the years 2015 and 2025. The comparison highlights three grammatical relations, revealing a notable semantic shift. In 2015, llama predominantly referred to the animal, as seen in collocations such as llama (was) lassoed. By contrast, in 2025, the term is primarily associated with large language models, as in fine-tune (a) Llama.

5. Conclusion

This paper presented the Trend Corpora – timestamped monitor corpora collected through a long-term, privately funded initiative across a wide range of languages (currently 25). These corpora are accessible via the Sketch Engine platform, which offers robust functionality for investigating linguistic trends (including neologisms) and language change over time.

Future plans include expanding existing Trend Corpora and extending the collection to Afrikaans, Amharic, Armenian, Azerbaijani, Georgian, Hausa, Igbo, Indonesian, Malay, Oromo, Tamil, Urdu, Uzbek, and Yoruba. These languages are the next candidate set to be added in the coming months.

Further efforts will focus on investigating why certain feeds have stopped supplying new data and on implementing automated methods to identify additional feeds within the downloaded content, thereby increasing the size and diversity of the corpora.

Figure 9: Sketch Engine Word Sketch – A three-dot menu for each collocate with links to the respective Concordance, Word Sketch, or Thesaurus

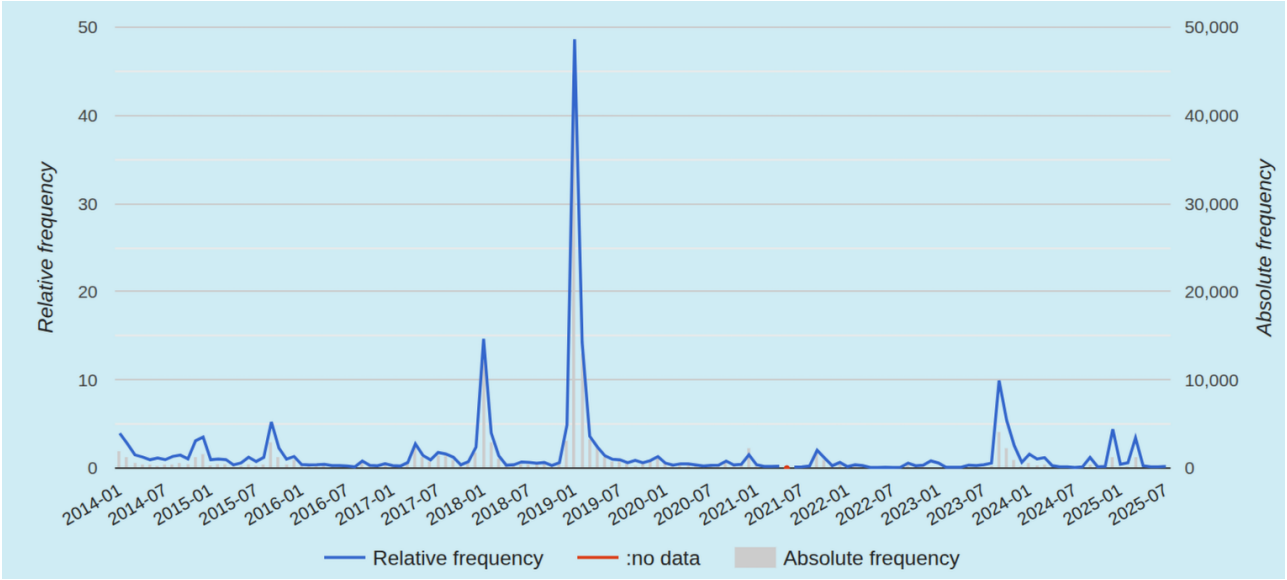

Figure 10: Sketch Engine Word Sketch – Timeline for the collocation *government shutdown*

Software

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and linguistic computing*, 25(4), pp. 447–464.

Davies, M. (2017). The new 4.3 billion word NOW corpus, with 4–5 million words of data added every day. In *The 9th International Corpus Linguistics Conference*, volume 2017. p. 2.

Herman, O. (2025). Automatic Detection of Word Sense Shift from Corpus Data. *Electronic lexicography in the 21st century. Proceedings of the eLex 2025 conference.*

Kilgarriff, A., Baisa, V., Busta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: Ten Years On. *Lexicography*, 1(1), pp. 7–36.

Kosem, I. (2022). Trendi - a Monitor Corpus of Slovene. In A. Klosa-Kückelhaus, S. Engelberg, C. Möhrs & P. Storjohann (eds.) *Dictionaries and Society. Proceedings of the XX EURALEX International Congress.* Mannheim: IDS-Verlag, pp. 230–239.

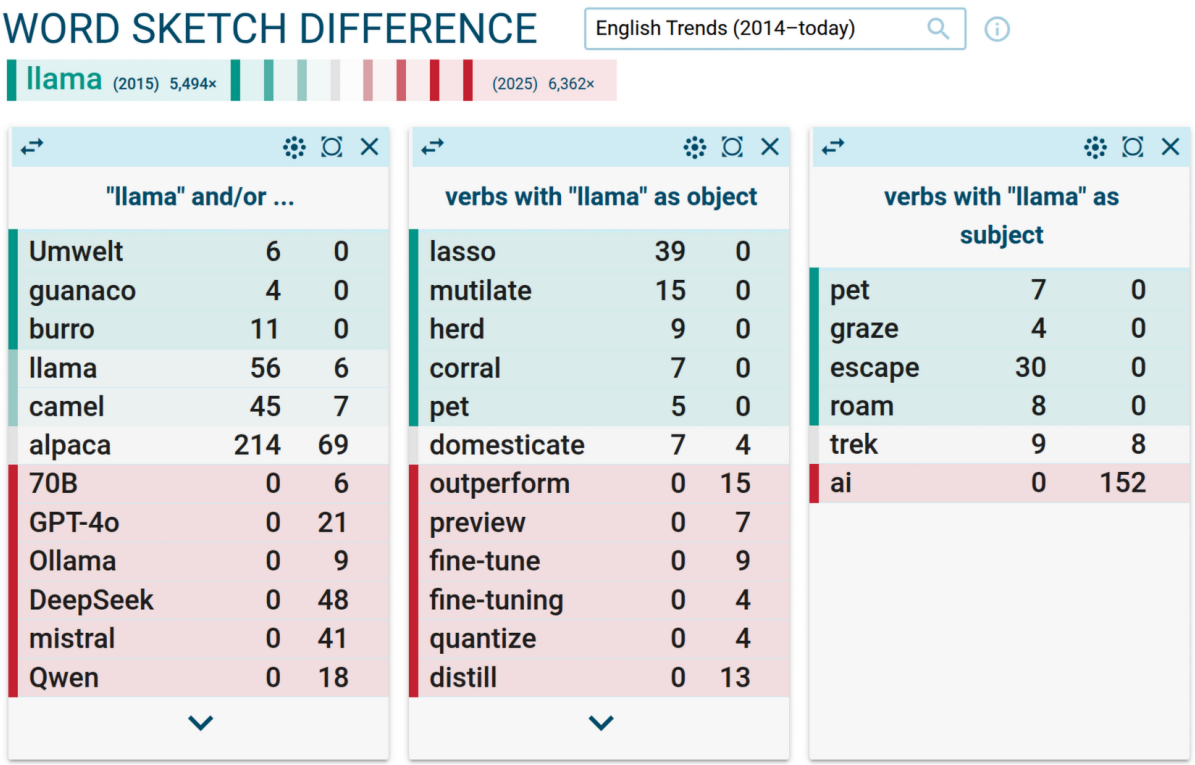

Figure 11: Sketch Engine Word Sketch Difference – comparison of occurrences of the word *llama* in subcorpora 2015 and 2025 by three grammatical relations; the figures in columns represent the number of occurrences in the respective subcorpus; the green background indicates 2015 collocates while the red one indicates 2025 collocates

Krek, S., Herman, O., Bušta, J., Jakubíček, M. & Novak, B. (2017). JSI Newsfeed corpus. In *The 9th International Corpus Linguistics Conference*. University of Birmingham.

Lin, Y., Michel, J.B., Aiden, E.L., Orwant, J., Brockman, W. & Petrov, S. (2012). Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*. Association for Computational Linguistics, pp. 169–174.

Ó Meachair, M., Bhreathnach, Ú. & Ó Cleircín, G. (2022). Introducing the National Corpus of Irish Project. In T. Fransen, W. Lamb & D. Prys (eds.) *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*. Marseille, France: European Language Resources Association, pp. 99–103. URL https://aclanthology.org/2022.cltw-1.14/.

Pomikálek, J. (2011). *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Masaryk University.

Suchomel, V., Kraus, J. et al. (2022). Semi-Manual Annotation of Topics and Genres in Web Corpora, The Cheap and Fast Way. In *RASLAN*. pp. 141–148.

Trampus, M. & Novak, B. (2013). Internals of an aggregated web news feed. In *15th Multiconference on Information Society*. pp. 221–224.