sciendo
DOI 10.2478/jazcas-2025-0014

WHEN DATA MEET TOOLS: USING THE MONITOR CORPUS FOR THE ANALYSIS OF LANGUAGE DEVELOPMENT

VÁCLAV CVRČEK[1] – MARTIN STLUKA[2] – KLÁRA PIVOŇKOVÁ[3]
[1]Department of Linguistics, Faculty of Arts, Charles University, Prague, Czech Republic (ORCID: 0000-0003-3977-2393)
[2]Department of Linguistics, Faculty of Arts, Charles University, Prague, Czech Republic (ORCID: 0000-0003-3294-3583)
[3]Department of Linguistics, Faculty of Arts, Charles University, Prague, Czech Republic & Department of Philosophy and History of Science, Faculty of Science, Charles University, Prague, Czech Republic (ORCID: 0009-0000-0990-7424)

CVRČEK, Václav – STLUKA, Martin – PIVOŇKOVÁ, Klára: When Data Meet Tools: Using the Monitor Corpus for the Analysis of Language Development. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 157 – 166.

Abstract: The aim of this paper is to introduce an infrastructure developed within the HiČKoK project to enable full-fledged corpus-based diachronic research of Czech. The individual sections of the paper present the components of this infrastructure, which links well-balanced, representative and annotated data with tailor-made tools for diachronic research. The forthcoming monitor corpus, covering the entire period of written Czech, along with its composition and annotation strategies, is briefly introduced. In the following sections, the potential of the application and its four modules—simple query, comparison, time-based associations, and diachronic collocations—are demonstrated through mini case studies. Combining large-scale data (as representative as possible) with a tool that enhances standard corpus functionalities, enriches them with a diachronic perspective, and enables result visualization makes diachronic research on language change more accessible and comprehensive.

Keywords: diachronic research, corpus querying, annotation, language change, monitor corpus, frequency, Czech

1 INTRODUCTION

The use of corpus linguistic methods for research on language change is accompanied by a number of specific challenges and issues. What has been addressed several times are the issues related to corpus compilation and annotation (Davidse and de Smet 2020). Compiling a diachronic corpus often requires sophisticated solutions for taming the variability of spelling (or even standardization of different writing systems); annotation systems have to be painstakingly adjusted to deal with the changes in morphology, syntax and lexicon in order to provide output that is both usable and adequate. A problem of its own, which in many cases cannot be solved satisfactorily, is the composition of the corpus and its representativeness, where we do not find the same spectrum of text types at different stages of language development.

The other group of issues is related to the tools and methods we use for exploring diachronic data. This has been addressed significantly less in the literature despite the fact that diachronic description calls for specific approaches and often requires specific measures for analyzing data.

In this paper, we would like to argue that for full-fledged corpus-based diachronic research, we need to address issues stemming from both the data and the tools for exploring them. This is because, in our view, a truly full-fledged infrastructure for working with diachronic data requires both well-constructed and annotated data and custom-tailored tools for examining them. This type of infrastructure for Czech is being developed within the HiČKoK project (Historie češtiny v korpusovém kontinuu 'History of Czech in the Corpus Continuum', see https://korpus.cz/hickok) which is supported by the Technology Agency of the Czech Republic within the Programme for Support of Applied Research and Innovation SIGMA. Its duration is from September 2023 to November 2026.

First, we will describe the Monitor Corpus that is being developed within the HiČKoK project (and its current version), briefly introduce the plan for its annotation, and in the second part we will focus more on the tools that we develop for effective work with this diachronic data.

2 THE MONITOR CORPUS OF CZECH

The aim of the HiČKoK (History of Czech in the Corpus Continuum) project is to create data, software and knowledge resources for the study of Czech throughout its history (13th–21st century).

One of the main objectives of the project is to create a Monitor Corpus of Czech that covers all eight centuries of the development of Czech in three main text types (fiction, non-fiction and journalism) where possible. It consists of all available diachronic data of the participating institutions (Czech National Corpus, Charles University and Institute of the Czech Language, Czech Academy of Sciences); part of the data that was not available in corpus format (covering the period of 1900–1990) was obtained in cooperation with the National Library of the Czech Republic. At the beginning of 2025, a working internal version of the corpus was created from all linguistically non-annotated texts by harmonizing data from each of the periods being processed (13th–15th centuries, 16th–18th centuries, 19th century and 20th century up to the present). The size of the corpus in its periods is summarized in Tab. 1.

Period	Tokens	Documents
13th–15th century	6 524 459	271
16th–18th century	2 809 406	197
19th century	18 233 175	999
20th–21st century	66 033 366	4 029
Total	93 600 406	5496

Tab. 1. Number of tokens and documents in the main periods of development of Czech and the total numbers for the whole corpus

The main goal, with respect to data compilation, is that the Monitor corpus of Czech should be able to represent the entire development of Czech in a single corpus, uniformly tokenized and annotated according to the latest standards. This is unique not only in the context of Czech, but also worldwide. Similar projects such as COHA (covering the period of 1920–2010), the Helsinki Corpus of English Texts (850–1710) or EEBO (containing over 25,000 books of various genres printed between 1475 and 1700) usually cover shorter periods of time, do not include contemporary language or are not designed as genre-balanced (fiction, non-fiction and journalism).

2.1 Corpus annotation

The methodology chosen for the project enables the corpus processing and annotation of Czech texts produced over eight centuries. The chosen processing approach takes into account the needs of the contemporary user while reflecting the linguistic evolution of Czech.

In attempting to cover eight centuries, care must be taken for comparability and consistency in annotation. For these reasons (and for reasons of cross-linguistic comparability), we have opted for the Universal Dependencies (UD; de Marneffe et al., 2021) framework, which serves as a de facto international annotation standard, to process the entire corpus. For these purposes, it was necessary to develop both a unified lemmatization system and to adapt synchronous tagging tools (cf. Zeman et al. 2023).

For this purpose, training datasets (etalons) with manually tagged and corrected texts were created that include samples of data from each period (see Tab. 1).

The corpus will be accessible by default via the KonText search interface for standard corpus querying, inspecting concordances or creating frequency distributions. Another output of the project will be freely available UD language models for annotating texts from different periods of Czech language development. The model for contemporary Czech is already available; models for older phases will be based on manually annotated (etalon) samples (for the earliest period up to the end of the 15th century, for Middle Czech from the 16th–18th centuries, and for the 19th century – approx. 100,000 tokens each).

In the following sections, we describe the possibilities of working with the Monitor Corpus within the Timeline Maker application. As has been pointed out, lemmatization and tagging are still in the process of development, so the demonstrations of working with Timeline Maker will be based on working with the beta version of the corpus, which so far contains only word forms and metadata.

3 TIMELINE MAKER

In contrast to synchronic research, which uses standard corpus tools (concordance, collocation, frequency distribution), diachronic research has an additional temporal dimension. At the same time, most of the standard tools for working with language corpora do not have the necessary functionality to capture and visualize phenomena during their change.

This issue is tackled by the Timeline Maker application, which is being developed within the HiČKoK project and is currently in its beta version. Its main advantage over standard corpus tools is the fact that while allowing for standard corpus querying it is designed to be able to work with diachronic data that contains information about the time (year) of creation of the text.

Timeline Maker is designed as a GUI on top of the KonText corpus manager (Machálek 2014) on the R/Shiny platform. Queries (in CQL format) that are entered into the application by the user are transformed into a series of queries to the KonText API so that they cover the entire timeline represented by the corpus (or a subpart of it selected by the user). The results obtained from the KonText API are then visualized via the Plotly library.

One of the key features of Timeline Maker is its ability to work with variable granularity of the timeline. This is a feature of crucial importance as in the case of diachronic data, which is relatively sparse, especially in the older phases of language development, it is often necessary to aggregate the results into larger units (decades, quarter-centuries or half-centuries) in order to be able to spot a development trend.

The application is divided into four modules, each representing a different type of query:

1. simple query: showing the frequency trend for a single phenomenon
2. comparison: represents the proportion of frequencies of two competing variants
3. companions: allows for detecting a similar (correlated) frequency trend of two phenomena, which might suggest an association between them
4. diachronic collocations: visualizes the change of the collocation profile of two words over time

In the following sections, each module will be described and exemplified on a selected phenomenon.

3.1 Frequency trend

The "simple query" module is designed to capture the frequency evolution of a given lexical or grammatical phenomenon. Its main goal is to plot the frequency of a given phenomenon in a graph (with the x-axis being the timeline), both in individual years and in aggregate form (in periods of 10, 25 or 50 years).

In addition to the query itself, which can be in the form of CQL and can contain standard regular expressions, the input form of this module allows the user to specify the time range in which the query will be evaluated.

For the "simple query" module, we show an example illustrating the gradual disappearance of the word *popravcě* ('judge/executioner'). In earlier periods of Czech, this was a polysemous lexeme related to legal topics, see ESSČ (*Elektronický slovník staré češtiny*, 2006–).

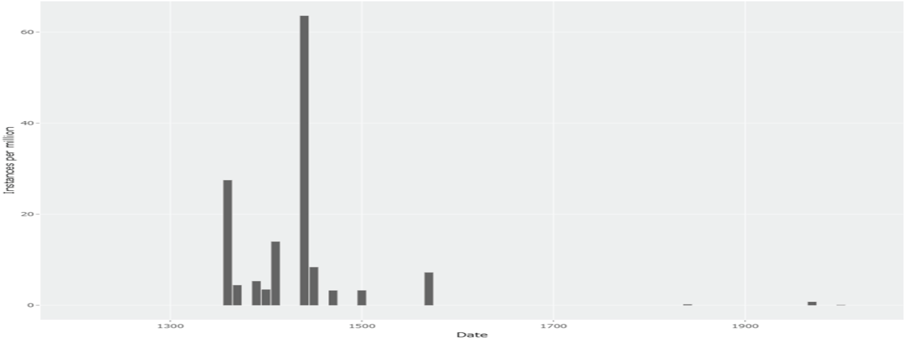


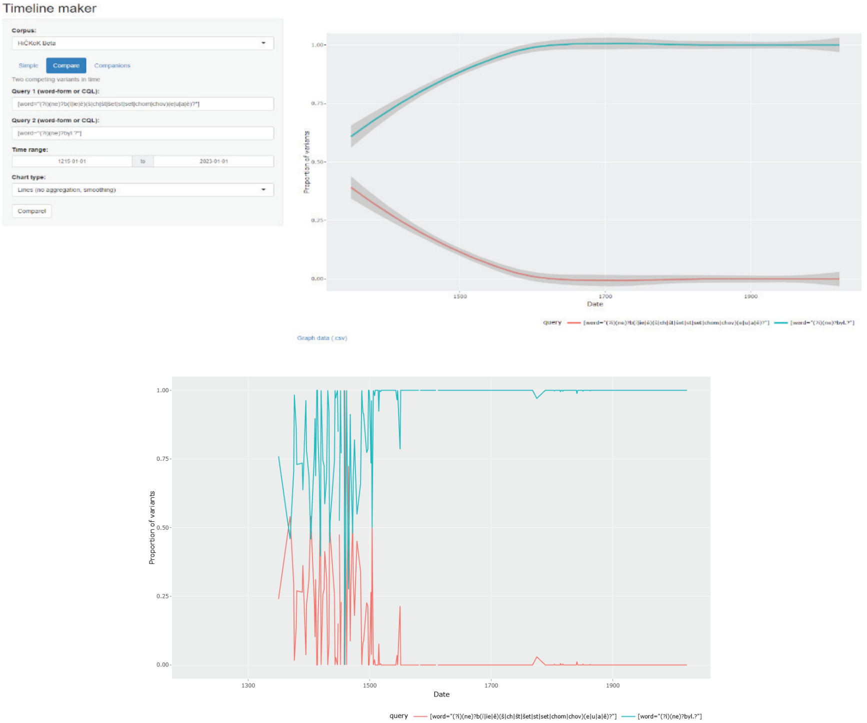
Fig. 1. Occurrence of the word *popravcě* 'judge/executioner'

The results of a search for *[Pp]oprávc.** by decade (shown in Fig. 1) indicate that the word *popravcě* ('judge/executioner') appeared consistently in the Old Czech period, namely from the mid-14th century to the end of the 15th century, while in the following periods, it is documented only in isolated instances. This suggests that from the end of the 15th century *popravcě* ceased to be part of the linguistic usage and was probably replaced by other terms (e.g. *soudce* 'judge', *kat* 'executioner').

3.2 Two competing variants

Similarly to the previous module, for comparing two competing variants, the user enters a CQL query (one for each variant) and selects the time window in which s/he wants to evaluate the query. This module is used to visualize the proportion of morphological, lexical or word-order variants over time. As in the previous case, this module offers the possibility of aggregating results by year, decade, quarter-century or half-century. In addition, it offers smoothing, which provides a better overview of the overall trend using the moving average method.

A suitable example to showcase the possibilities of this could be the gradual disappearance of the simple past tenses in Czech (aorist and imperfect), specifically for the verb *být* ('to be'). Since the Monitor corpus is not yet annotated, a more complex query was required: `[word="(?i)(ne)?b(í|ie|ě)(š|ch|št|šet|st|set|chom|chov)(e|u|a|ě)?"]` as the first variant versus `[word="(?i)(ne)?byl.?"]` as the second. For the sake of clarity, we have chosen a visualization without any aggregation and the method with smoothing with local polynomial regression and 95% confidence intervals (upper chart of Fig. 2) and without it (lower chart of Fig. 2). The upper-left section of Fig. 2 displays the application's input form for this module.



Timeline maker

Corpus: HiČKoK Beta

Two competing variants in time

Query 1 (word-form or CQL): [word="(?i)(ne)?b(í|ie|ě)(š|ch|št|šet|st|set|chom|chov)(e|u|a|ě)?"]

Query 2 (word-form or CQL): [word="(?i)(ne)?byl.?"]

Time range: 1215-01-01 to 2023-01-01

Chart type: Lines (no aggregation, smoothing)

Fig. 2. Decline of the simple past tenses in Czech – horizontal axis represents years, vertical axis shows percentage of competing variants

The results (Fig. 2) indicate that the decline of the simple past tenses in Czech has been continuous since the time they were consistently documented in writing, approximately from the 13th century. This process, closely linked to the development of the aspectual system in Czech, was completed in the written language by the early 16th century and likely even earlier in the spoken language (to read more about the decline of simple past tenses in Czech, see, e.g. Kosek 2017). In case of the verb *být*, which is highly frequent (as its aorist and imperfect forms were also used in periphrastic constructions, e.g. *bieše dělal* 'he was doing'), it can be assumed that these past tense forms persisted longer in written texts than in case of less common or functionally limited verbs.

3.3 Time-based associations

The third module, called Companions, offers a new insight on simultaneous changes in language. It tracks the frequency development of two phenomena in time. Synchronization strength (peaks and valleys of the frequency trend of both phenomena under examination) can be measured by cross-correlation (used in signal processing, for example). It compares the frequency development curves of two words or other phenomena over time and evaluates the similarity of their shapes. In addition to the correlation coefficient (r), it also calculates the lag between curves, by which one phenomenon is delayed in its development relative to another.

Companions, as a method, was originally developed to measure how two words (or other phenomena), triggered by some real-life event, start occurring and gain or lose frequency in the same time slots; in this sense the two words become "companions" in discourse/language (Cvrček and Fidler 2024).

The query input for this module is the same as for the variant comparison. In addition, the user can choose to see both trends in one graph or in graphs below each other. Beyond the visualization, this module also shows the result of the cross-correlation measure (r), which represents the degree of association between the observed phenomena.

The intent of this module is not to explore collocations or other phenomena based on syntagmatic relations, which we could easily inspect by concordancing or by other corpus tools; it is designed to examine the correlation of independent words usually linked by some real-life change. The identification of suitable 'companion' candidates is thus not straightforward. This requires at least a basic historical and socio-cultural awareness of the period under study and an idea of what might be worth looking for. As an example, we chose the words (forms of) *válka* 'war' and *mor* 'plague' to analyze the frequency of their occurrence in the period from 1400 to 1630.

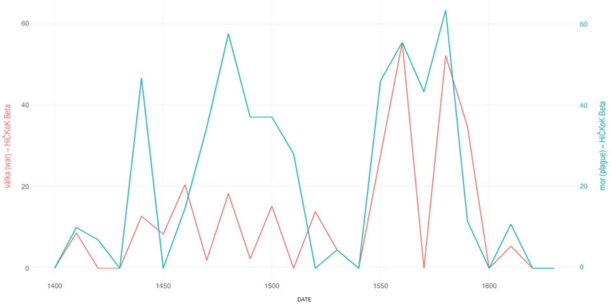


Fig. 3. Time-based association between the words *válka* 'war' and *mor* 'plague'

Both words are presented together for comparison in Fig. 3 (with the respective y-axes on either side), which demonstrates a relatively high degree of time-based association between them (r=0.62). While we can not be 100% sure that these words are true companions (i.e., they are not just randomly correlated), it strongly suggests that *válka* (‘war’) and *mor* (‘plague’) were not merely coincidentally associated during the examined period. By the end of the 17th century, the degree of association had already decreased, indicating that wars were likely no longer as commonly accompanied by plagues as they had been in earlier periods.

3.4 Semantic shift

The last module is designed for detecting the semantic shift of words. For this purpose it uses the comparison of the collocation profiles of a given word in different time periods. Collocation profiles are created for the word under examination at individual time intervals according to the user’s specification (5, 10, 20... years). The user can further specify the association measure that will be used to identify the most significant collocations (log-dice, log-likelihood, MI score, t-score), the context window in which collocations will be evaluated and the minimum collocation frequency (to filter out rare phenomena). Another option in the settings allows ignoring collocates that appear in isolation (in one time period only).

The results for each period are then visualized using a special flow-chart that allows to track the changes in the collocation profile and infer the change in meaning of a word (while maintaining the same form).

To demonstrate the potential of this module for detecting semantic shifts, we chose an example of a word that has undergone substantial semantic change over several centuries. Although there is an abundance of such words, in order to satisfactorily track their semantic change with this application, a sufficient frequency over all the periods studied is required, as this is the only way to obtain statistically reliable results. More than satisfactory results can be presented on the word *páteř* (which in modern Czech means ‘spine’) using the following query settings: time interval 1215–2023, aggregation period 25 years, collocation (association) metric log-dice, context window -3 +3, and a minimum collocation frequency of five.

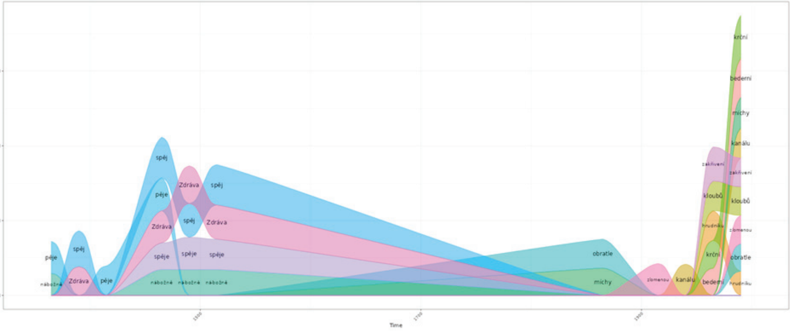


Fig. 4. Semantic shift of the word *páteř* ‘spine’ – the width of the strand for each collocate represents its association strength measured by logDice.

The meaning of the word *páteř* in modern Czech, i.e. the basic part of the skeleton of vertebrae, can be found as early as the 15th century, but at that time it played only a marginal role (see also ESSČ). As can be clearly seen from the collocation profile (Fig. 4), the word *páteř* had a different dominant meaning in Old Czech, that of *Otčenáš* 'Lord's Prayer' (note: Interestingly, the Old Czech dictionary documents another meaning of the word *páteř* in Old Czech – besides 'spine' and 'Lord's Prayer' – namely, 'rosary', which is also related to prayer, as it refers to a string of beads used for counting prayers, particularly repetitions of the Lord's Prayer). Typical collocations for the word *páteř* in Old Czech are words semantically related to prayer: *nábožně* 'piously', *(s)pěti* 'recite', and *Zdráva* 'Hail' (referring to the prayer of the Hail Mary). From around the 18th century, the results show a semantic shift. Collocates from this period are semantically related to the vertebral skeleton, such as *obratel* 'vertebra', *mícha* 'spinal cord', *kloub* 'joint', *bederní* 'lumbar', and *krční* 'cervical'. We can clearly observe a diachronic shift in which the meaning of 'spine' becomes dominant.

4 CONCLUSION

In this paper, we have attempted to illustrate that in order to effectively use corpus methods for diachronic research, two conditions need to be met: 1. corpus data that is representative of the time period, with consistent processing, annotation and metadata, 2. dedicated tools that are capable of evaluating phenomena on a timeline.

For the diachronic study of Czech, the first condition should be met prospectively by the Monitor Corpus of Czech, which is the largest achievement in this field so far, and which fulfills the ambition of a complete coverage of the timeline of Czech language development, uniform annotation and processing.

The second condition is prospectively met by the Timeline Maker application, which is specifically designed with the intent to analyze data in a diachronic perspective. It complements standard corpus tools with visualization capabilities that help interpret developmental phenomena in the language.

ACKNOWLEDGEMENTS

The HiČKoK project (History of Czech in the Corpus Continuum, see https://korpus.cz/hickok) is supported by the Technology Agency of the Czech Republic within the Programme for Support of Applied Research and Innovation SIGMA (no. TQ01000072).

This work has been supported by Charles University Research Centre program No. 24/SSH/009.

References

COHA (Corpus of Historical American English). (n.d.). English Corpora. Accessible at: https://www.english-corpora.org/coha/ [29/03/2025].

Cvrček, V., and Fidler, M. (2024). From News to Disinformation: Unpacking a Parasitic Discursive Practice of Czech Pro-Kremlin Media. Scando-Slavica, 70(1), pp. 32–54. Accessible at: https://doi.org/10.1080/00806765.2024.2317374.

Davidse, K., and De Smet, H. (2020). Diachronic corpora. In: M. Paquot – S. Th. Gries (eds.): A practical handbook of corpus linguistics, pp. 211–233. Springer International Publishing. Accessible at: https://doi.org/10.1007/978-3-030-46216-1_10.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal dependencies. Computational Linguistics, 47(2), pp. 255–308. Accessible at: https://doi.org/10.1162/coli_a_00402.

EEBO (Early English Books Online). (n.d.). English Corpora. Accessible at: https://www.english-corpora.org/eebo/ [29/3/2025].

Elektronický slovník staré češtiny [online]. (2006–) Praha: Ústav pro jazyk český AV ČR, v. v. i., oddělení vývoje jazyka [cit. 20/06/2020]. Accessible at: http://vokabular.ujc.cas.cz.

Kosek, P. (2017). IMPERFEKTUM. In: P. Karlík – M. Nekula – J. Pleskalová (eds.): CzechEncy – Nový encyklopedický slovník češtiny. Accessible at: https://www.czechency.org/slovnik/IMPERFEKTUM [last accessed 29/03/2025].

Machálek, T. (2014). KonText – aplikace pro práci s jazykovými korpusy [Cs]. FF UK. Accessible at: https://kontext.korpus.cz.

Rissanen, M., Kytö, M., and Heikkonen, K. (eds.). (1991). The Helsinki Corpus of English Texts: Diachronic and Dialectal. University of Helsinki.

Zeman, D., Kosek, P., Březina, M., and Pergler, J. (2023). Morphosyntactic annotation in universal dependencies for old czech. Jazykovedný časopis/Journal of Linguistics, 74(1), pp. 214–222.