Citation: Změlík, R. (2024). Computational literary studies in the context of contemporary Czech literary studies – especially the concept of the Corpus of Czech Novels (J. Línek, Trans.). *Colloquia Humanistica, 2024*(13), Article 3107. https://doi.org/10.11649/ch.3107

Colloquia Humanistica 13 (2024)
Digital Studies of Culture in Central Europe
Article No. 3107
https://doi.org/10.11649/ch.3107

COLLOQUIA HUMANISTICA

Richard Změlík

Faculty of Art
Palacký University in Olomouc, Olomouc
https://orcid.org/0000-0002-5414-4574
richard.zmelik@upol.cz

Computational Literary Studies in the Context of Contemporary Czech Literary Studies – Especially the Concept of the Corpus of Czech Novels[1]

Abstract

The paper discusses some issues related to the field called computational literary studies (CLS), which belongs to the digital humanities. The introduction deals mainly with the Czechoslovak literary and partly linguistic tradition that preceded this discipline. The author then introduces Czech projects that represent specialised literary corpora, in particular focusing on the emerging corpus and web application of the Czech Corpus of Prose, describing its overall structure, individual parameters and specific functions.

Keywords: computational literary studies, quantitative and cartographic models, structuralism, Czech literary studies, corpus linguistics, corpus in literary studies.

[1] The preparation and publication of this paper was made possible by financial support provided by the Faculty of Arts of Palacký University in Olomouc in the years 2022–2024 from the Fund for Support of Scientific Activities FPVC2022/20.

This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits redistribution, commercial and non-commercial, provided that the article is properly cited. © The Author(s) 2024.
Publisher: Institute of Slavic Studies, Polish Academy of Sciences
Editor-in-chief: Jolanta Sujecka
Conception and academic editing: Robert Kulmiński and Ondřej Daniel

Introduction

The unprecedented growth of digital technologies, which surround us today at almost every step and have become an inherent part of our lives, both public and private, has naturally permeated the humanities and social sciences. This is not surprising; there was a similar situation more than half a century ago, when the development of cybernetics and information theory after World War II influenced many disciplines of humanities-oriented research: linguistics, aesthetics, and literary studies. Any such "breakthrough" in the social sciences or humanities entails the necessary need for critical reflection and historical contextualisation, which arise not only from the need to understand their historical situation, but also their theoretical anchorage and the potential development that results from interdisciplinary overlaps. In scientific disciplines, no new impulse arises "out of thin air"; it is always preceded by a certain tradition that stems from the epistemology of the discipline. However, the current socio-cultural situation has a significant influence here, which – as in this case – is significantly influenced by the new possibilities that modern technology presents us with today. In this article we intend to focus on the issue of computational literary studies (CLS) in the Czech literary context, which can be considered part of the digital humanities (Debnár & Gogora, 2019, p. 15). We will present the current situation in this discipline and then introduce the emerging project of a corpus of Czech prose, which aims to fill a gap in this field.[2]

The Position of Computational Literary Studies in the Czech Literary Context Today

Although our focus here is exclusively on computational literary studies, we believe that they also belong to the context of the digital humanities, even though this term is changing greatly and taking on new connotations. Thus, let us start from a general definition of what the digital humanities

[2] It is not possible to discuss here all existing digital humanities examples in the Czech Republic. This task would necessarily require a separate study. Of the many projects that have been carried out in the Czech Republic, we will therefore mention two that are closest to literary science: *DL4DH*, which focuses on "machine recognition of biblical quotations in a corpus of selected First Republic periodicals from 1925–1939" (https://dl4dh.nkp.cz/; *DL4DH*, n.d.), and *Manuscriptorium* (https://www.manuscriptorium.com/cs; *Manuscriptorium*, n.d.), a project focusing on the digitisation of Old Czech manuscripts using XML transcription.

(henceforth DH) are and are not. In Anne Burdick et al.'s *Digital_Humanities*, we learn that "Digital humanities projects can be described by sketching their structure at several levels. These begin with basic computation (programming, processing, protocols) and extend through the levels of organisation and output that form the basis of most users' experience (interface, devices, networks) ... The second level involves processing in ways that conform to computational capacities, and these were explored in the first generation of digital scholarship in stylometrics, concordance development, and indexing. This processing activity takes advantage of the ability of computers to automate certain tasks useful in answering the sorts of research questions that were initially being asked by humanities scholars. ... Both computational foundations and processing activities have endured, but other platforms, tools, and infrastructures have also developed to support curation, analysis, editing, and modelling. These depend upon the basic building blocks of digital activity: digitisation, classification, description and metadata, organisation, and navigation. Designing and building digital projects depend on knowledge of these fundamentals and on a nuanced understanding of the networked environments in which the projects will develop and variously reside" (Burdick et al., 2012, pp. 16–17).

In Mark Debnár and Andrej Gogora's text *Digitálne trendy v súčasných humanitných vedách* [Digital Trends in Contemporary Humanities] (2019), we learn more about the definition of the digital humanities, which the authors define as "a field of interdisciplinary research in the humanities that is carried out through computational processes, creates and uses digital resources, extends traditional methodological practices and scholarly outputs, includes the training of digital scholars, and supports a collaborative community" (Debnár & Gogora, 2019, p. 15). Specifically, the authors then list the basic characteristics of DH research, which include the aforementioned interdisciplinarity, digital tools and resources, the relevant methodology based on the functional connection of digital tools (environment) with, say, the traditional requirements placed on social science or humanities research, education in the use of digital technologies, and the digital research environment. These authors do not understand DH as mere digitisation, trivial use of digital tools, individual research activity, digital culture and IT service (Debnár & Gogora, 2019, pp. 26–28).[3]

[3] It would be possible to disagree with some points. This is particularly the case under the heading "individual research activities". Although DH is generally interdisciplinary by nature, it does not exclude individual research activities, especially if the author is able to

Although we may argue about the general definition of DH, one thing remains a given: DH defines those social science and humanities disciplines that actively and, above all, functionally use digital technologies and their tools, which modify the methodology of these disciplines and enrich their results with new outcomes. As we will soon try to show with concrete examples and then with the specific project of the Czech prose corpus, CLS projects in the Czech Republic not only fully meet these basic definitions of the digital humanities, but also strive not to remain exclusively within the closed framework of private and externally inaccessible academic research by sharing their data and models publicly through web platforms.

The current state of the digital humanities in the Czech social sciences and humanities is best and comprehensively represented by the publication *Digital Turn in the Czech Humanities and Social Sciences* (2022), which presents research topics from different disciplines that actively and functionally use digital tools and algorithms that are embedded in the theoretical frameworks of their respective disciplines. From the content of the publication, it is possible to get a representative idea of how DH is developing in the Czech environment. It seems that the diversity of the field proves that the digital humanities are quite successful in the Czech Republic. The real situation is, however, somewhat more complex, as evidenced mainly by the current setting of research output evaluation by the Czech Ministry of Education. On the one hand, we can see a diverse use of different DH approaches and methodologies in the social sciences and humanities, but on the other, such research faces difficulties in the Czech environment due to the setup of the evaluation of the results of scientific work. The latter does not yet take into account standard DH research outputs such as datasets and databases, general and special language corpora, machine learning models, web and desktop applications, and many other types of DH outputs.[4]

In this context it might be interesting to mention the article "No 'Making', Not Now: Decolonizing Digital Humanities in South Asia" (2019), in which authors Dibyadyuti Roy and Nirmala Menon describe the situation

provide the technical aspects of the research. I am thinking in particular of the creation of scripts or digital (web) environments, etc.

[4] In 2023, the Czech Association for Digital Humanities (CzADH) published an open call to the Ministry of Education of the Czech Republic, the *Czech Association for Digital Humanities Opinion on the Evaluation and Recognition of Research Results*, which warns of the insufficient incorporation of specific DH research outputs in the methodology of evaluation of scientific work (*Stanovisko České asociace pro digitální humanitní vědy k hodnocení a uznávání výzkumných výsledků*, 2023).

of the digital humanities in India and South Asia. While that state of affairs is, of course, not comparable to the situation in the Czech Republic, a number of things are remarkably similar. The authors state that despite a number of sub-research projects that actively use digital technologies, the discipline has not fared well in terms of university departments in India: "We consider some of the reasons for the resistance to institutionalising DH within the larger Indian university systems and conclude that the causes are historical but perhaps also arise from the embedded rigidity and regimentation of university curricula across disciplines. DH, as defined by its purported interdisciplinary or transdisciplinary character, proves a tough sell within Indian university ecosystems, often constrained by their colonial origins and the limitations of a postcolonial social imaginary" (Roy & Menon, 2022, p. 187). Yet, as they further note, "While digital humanities as a subject, program, or course was rare, the use of technology in various economics, sociology, literature, and other humanities departments was quite prolific" (Roy & Menon, 2022, p. 189). Although the situation in India and the Czech Republic is incommensurable, when compared to the situation elsewhere in the world (especially in English-speaking and German-speaking regions), we do see some framework analogies.

What applies to DH in general, of course, also applies to CLS, in particular the current inability to include certain research outputs in the system of scientific evaluation of publication outputs. Nevertheless, the situation in the Czech Republic in this field (CLS) is improving year by year, which is evident not only in primary research (see below), but also in teaching at universities. For example, at the Faculty of Arts of the Palacký University in Olomouc, a doctoral programme in the digital humanities has been established, which is mainly focused on natural language processing (NLP) (*Digital humanities*, n.d.).[5] Although this is not directly a CLS field, there are still numerous overlapping areas, which naturally result from some similar methodological approaches at the level of NLP.

[5] Available at: https://kol.upol.cz/studium/moduly/digital-humanities/ It is necessary to mention here that NLP is taking off much faster in linguistics, which is natural in view of the tradition of corpus linguistics that began in the Czech Republic with the establishment of the Institute of the Czech National Corpus in 1994 at the Faculty of Arts of Charles University. Today it is probably one of the most sophisticated corpus platforms ever. In this context, we should also mention the Institute of Formal and Applied Linguistics at Charles University, where a number of useful NLP tools are being developed that can be used in the context of CLS. There is also an optional *Digital Humanities for Literary Science* seminar at the Palacký University in Olomouc, aimed at teaching the basics of programming in Python and focused on NLP.

CLS and the Tradition of Czechoslovak Literary Science

Quantitative and statistical methods of analysing literary texts form an integral part of CLS. Some selected methods of quantification and statistics can already be found in the environment of interwar Czechoslovak structuralism.[6] First of all, these are the versological works of Jan Mukařovský and Roman Jakobson, in which the researchers quantified the metrics of selected Czech poems (see Jakobson, 1995a, 1995b; Mukařovský, 1982a, 1982b). Although this methodological step can be considered progressive from the point of view of the contemporary situation of Czech and international literary scholarship, Lubomír Doležel did not hesitate to label it as pre-statistic almost 40 years later (see Doležel, 1965, p. 104). What led Doležel, otherwise a prominent heir of the Prague structuralist school, to such a conclusion? It was primarily the way in which Czechoslovak structuralists worked with quantitative models. According to Doležel, their use was illustrative and additive rather than functional. On the other hand, Doležel mentions in his work the need for active involvement of quantitative and statistical models as well as cybernetic procedures in concrete literary work, which he also subsequently demonstrates with the example of distinguishing the stylistic characteristics of poetic texts.

The 1960s in particular were a period when innovative methodological approaches inspired by information theory and cybernetics began to be promoted not only in linguistics but also in literary science in Czechoslovakia. In addition to Doležel, in the 1960s and 1970s these were explored by Jiří Levý (1971) and Eduard Petrů, who, in his work entitled *Exaktní metody v literárněvědné práci* [Exact Methods in Literary Science] (Petrů, 1968), proposed, among other things, a punched-label system for cataloguing Old Czech literary monuments. This tradition was continued in the 1980s by Pavel Vašák, whose contributions in this field include the incorporation of these methods into textological research (Vašák, 1980), and in the 1990s by Miroslav Červenka and Květa Sgallová

[6] In 1926 Vilém Mathesius founded the Prague Linguistic Circle at Charles University, which developed structural approaches in linguistics, literary science, aesthetics and theatre studies. Russian scholars also became part of the circle, including, for example, Roman Jakobson, who had previously belonged to the Moscow Linguistic Circle, which, together with the St. Petersburg OPOYAZ, participated in the proclamation of formalist methods in linguistics and literary science in Russia.

(Červenka & Sgallová, 1995), who applied quantitative and statistical methods in Czech versology. These scholars, who were closely tied to the Czechoslovak structuralist tradition, managed to develop the impulses arising from information theory and cybernetics in an original way in Czechoslovak and Czech literary science. The fundamental prerequisite for the development of these approaches was, of course, Czechoslovak structuralism, which, following de Saussure's linguistics, emphasised the autonomy and complexity of the literary sign, which is determined by its internal organisation, the way in which the sub-units are arranged as well as the way in which they relate to each other. The idea of scientific exactness, which was at the basis of Czechoslovak structuralism and which was also significantly inspired by Russian formalism, could thus follow the new post-war tendencies in the social sciences and humanities inspired by the aforementioned information theory and cybernetics. Here, the concepts of code, coding, system, structure, and quantity as one of the structural attributes of elements in an ordered system, formed a common intersection.

The post-Soviet situation affected Czech literary science with similarly dramatic changes, as it was de facto reflected in all areas of social and cultural life. The main effort was to balance the theoretical and methodological rupture between Czechoslovak and foreign, especially Anglosphere, literary science and aesthetics. Thus, narratology, semiotics, deconstruction, the theory of fictional worlds, phenomenology, or, in recent years, more thematically oriented interdisciplinary research such as gender studies, feminist literary criticism, queer or animal studies, etc., have come to the fore in the Czech Republic. In the last few years, Czech literary scholarship has also begun to talk about the potential of the digital humanities. It is even possible to identify a certain milestone, namely 2017, when the first public expert debate on the digital humanities in literary studies was held, organised by the Institute for Czech Literature of the Czech Academy of Sciences (Ústav pro českou literaturu AV ČR, 2020). However, even before then, some of the first projects and outputs that today are referred to as computational literary studies, were already being produced; namely, Petr Plecháč's project *The Corpus of Czech Verse* (see below) (Plecháč, n.d.) and Richard Změlík's monograph *Kvantitativně-korpusová analýza a literární věda* [Quantitative Corpus Analysis and Literary Studies] (Změlík, 2015).

Examples of Computational Literary Studies in Czech Literary Science

Currently, several partial projects in the field of computational literary studies are being implemented in the Czech Republic, at universities or at the Czech Academy of Sciences (CAS). In recent years, the Institute of Czech Literature of the CAS created a representative project, the *Corpus of Czech Verse*, which includes Czech poetry of the 19th and early 20th centuries and provides tools for quantitative versological analysis of meter, rhyme and euphony. The same department also produces the *Corpus of Contemporary Poetry* (Piorecký & Škrabal, n.d.), which was made available in 2022 and focuses on Czech poetry published since 1989. The project partly uses the versological analysis tools of the previous project; the corpus is otherwise part of the digital infrastructure of the *Czech National Corpus* (*Český národní korpus*, n.d.). The last major project of the aforementioned department is the *Czech Literary Bibliography* (*Česká literární bibliografie*, n.d.), which collects and structures bibliographic databases from the beginning of the 19th century to the present day, and its activities continue the long bibliographic tradition of the Institute for Czech Literature of the CAS. These are actually all the larger projects of this kind in Czech literary studies, although it should be mentioned that further projects are currently being prepared, one of which will be presented in the next section.

The same is true in terms of publishing. Some of Franco Moretti's works have been translated into Czech, namely the book *Graphs, Maps, Trees: Abstract Models of Literary History*, the Czech translation of which was published in 2015 (Moretti, 2015), and the study *Network Theory and Synopsis Analysis*, published in Czech in 2019 (Moretti, 2019), and also the book *Digital_Humanities* (Czech translation 2019; Burdicková et al., 2019) by Anne Burdick, Johanna Drucker, Peter Lunenfeld, Todd Presner and Jeffrey Schnapp. Yet so far, unlike foreign research, this area is relatively poorly covered. Among domestic works dealing with the digital humanities, specifically in literary studies, we can mention mainly works in the field of quantitative versology. Its authors are members of the versology team at the Institute for Czech Literature of the CAS (ÚČL AV ČR). There are also two monographs by the author of this study (see Změlík, 2015, 2019),[7] and the representative

[7] For the sake of completeness, let us add that at the time of writing this article, a monothematic issue of the journal *Česká literature* [*Czech Literature*] devoted to the digital humanities in literary science was in print.

collective monograph *Digital Turn in Czech Humanities and Social Sciences* (2022) already mentioned earlier.

Since the first public expert debate on the digital humanities in literary studies in the Czech Republic, which took place at the Institute of Literary Studies of the CAS in 2017, the situation has improved significantly and Czech literary scholars are gradually accepting this type of research, as evidenced by the monothematic sixth issue of the journal *Česká literatura* ("Kvantitativní a statistické metody," 2022),[8] which was devoted, among other things, to issues of quantification, statistics, stylometry, cartographic models and digital literature.[9]

Corpus of Czech Prose

Whereas two significant corpora of poetic texts already exist in the Czech literary scholarly environment (see above), a digital corpus of Czech passive prose is missing for the time being, or is currently being implemented at the Czech Department of the Palacký University in Olomouc. In the following section we would like to present its main parameters as well as the possibilities of its mining.

The initial intention for building the corpus of Czech prose (officially called *Literary Cartographic and Quantitative Models of Czech Novels from the 19th to 21st Century*; Změlík, n.d.[10]) was to collect texts of Czech prose from the 19th to 21st centuries that explicitly thematise the Prague environment, with the aim of creating, on the one hand, cartographic models of fictional Prague topography and, on the other, a set of quantitative models that reflect selected features of narrative artistic texts. Gradually, however, other texts are added to these texts, which no longer necessarily meet the above-mentioned criterion. The reason for this extension is the gradual addition of several corpus functionalities that are not limited exclusively to questions related to mapping or quantifying selected segments of narrative text, but allow partial and total searches oriented towards selected quantitative, statistical criteria (see below). This extension turns the project into a tool that, in the end, is aimed primarily at providing a functional link between literary science and linguistics.

[8] Available at: https://asjournals.lib.cas.cz/ceskaliteratura/archiv;pid=uid:74b544c8-9bed-4166-8bb7-c4a1a9893bd5

[9] Karel Piorecký has long been working on this issue in Czech literary studies (see Piorecký, 2016).

[10] Available at: https://korpusprozy.com. Although the year 2020 is mentioned on the website, this is the year of the launch of the web platform. The preparatory work is older.

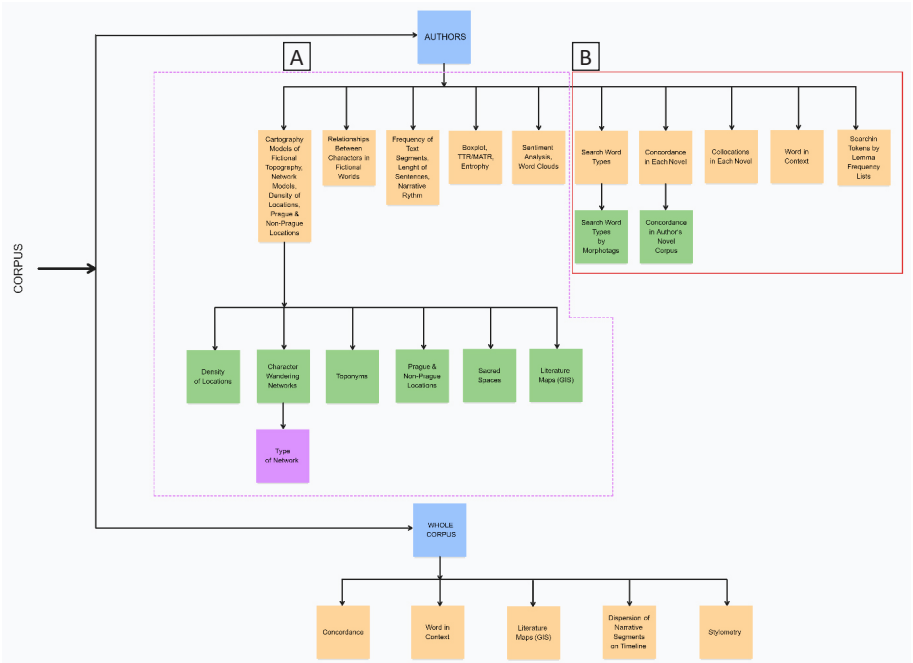


Figure 1. Map of the web application.*

If we look at Fig. 1, this intention corresponds to two interconnected parts of the corpus (A – dotted pink frame, B – solid red frame), one of which includes cartographic models, network models and quantitative models of textual segments of the artistic narrative, while the other is dedicated to features for searching the linguistic corpus of literary works according to selected criteria, i.e. concordance, bigram collocations, sentiment analysis, searching for lemmas, tokens and their frequencies, searching according to complex morphological tags or selected words in close context (see below). As the diagram also shows, the corpus can be searched not only with respect to individual authors and their works, but also as a whole, and concordance and close context word search functions are available in this module for now.

The prerequisite not only for this project, but de facto for any language corpus, is not only data heuristics, but also data processing, i.e. mainly data

* All illustrations in the paper were created by the author using tools described in the paper.

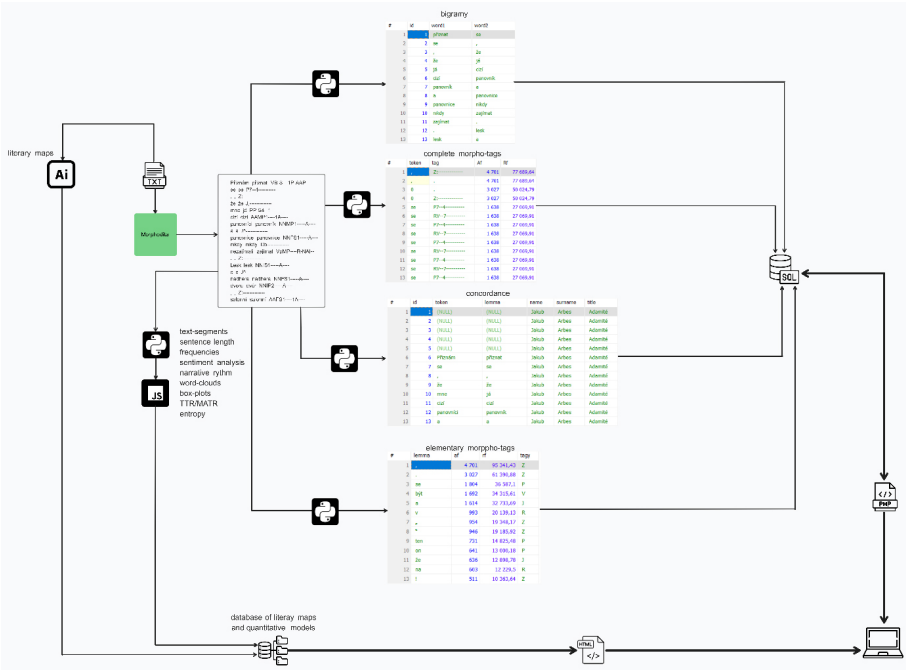


Figure 2. Schematic depiction of data processing and incorporation into the prose corpus.

segmentation and annotation. Fig. 2 shows the basic scheme of how the text data was processed and incorporated into the Czech prose corpus.

As can be seen from the diagram, the plain text is initially lemmatised using the MorphoDita language tool,[11] which segments the text into tokens, type and morphological tags.[12] Subsequently, sub-tables are created which are designed to form the data structure of the corpus, depending on the search criteria. For example, as can be seen in Fig. 2, the concordance table contains data in six columns (ID, token, lemma, name, surname, title). Tables for individual works and specific authors are stored in databases

[11] The application was developed by Jana Straková, Milan Straka and Jan Hajič at the Institute of Formal and Applied Linguistics of Charles University. Available at: https://ufal.mff.cuni.cz/morphodita

[12] This is a 16-character morphological tag that is assigned to each token. Depending on the word-species specification of the token, the tag contains additional relevant information about the grammatical categories.

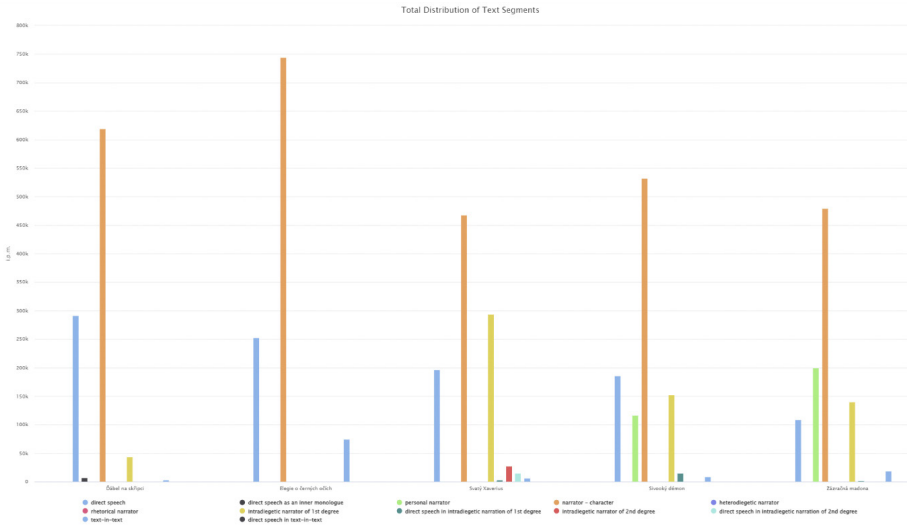


Figure 3. Frequency loading of selected text segments of the artistic narrative in five novels by Jakub Arbes: *The Devil on the Scissors*, *Elegy of the Black Eyes*, *Saint Xaverius*, *The Grey-Eyed Demon* and *The Miraculous Madonna.*

in the MySQL environment, from where data is extracted based on user queries. In addition to these data, the server also hosts a database of cartographic models of fictional Prague topographies, which were developed in vector software and which contain overlay layers carrying partial topographic-textual information.[13]

The quantitative models are represented in the corpus by both frequency load graphs of text segments (a specific selection of these is shown in Fig. 3) and graphs showing narrative rhythm, i.e. the frequency alternation of different narrative utterances (see Fig. 4), or charts for measuring so-called sentiment analysis (see Fig. 5). In this case, two different methods were chosen. The first one uses the available lexicon of Czech words, which are classified as either positive or negative. The lexicon was compiled at the Institute of Formal and Applied Linguistics.[14] The second method also uses the dictionary data type (Json), but this is processed in a slightly different way. Here, individual words are assigned to a general concept denoting an

[13] A detailed description is part of another study published here under the title "Literary-Cartographic Models in the Corpus of Czech Prose".

[14] Available at: https://ufal.mff.cuni.cz/sentiment/data

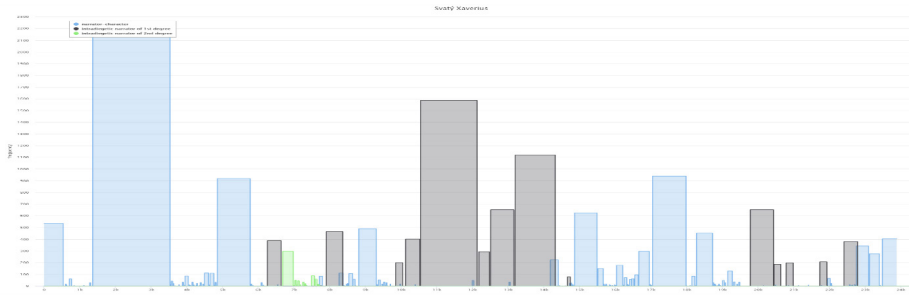


Figure 4. Narrative rhythm in the prose *Svatý Xaverius* [Saint Xavier] by Jakub Arbes.

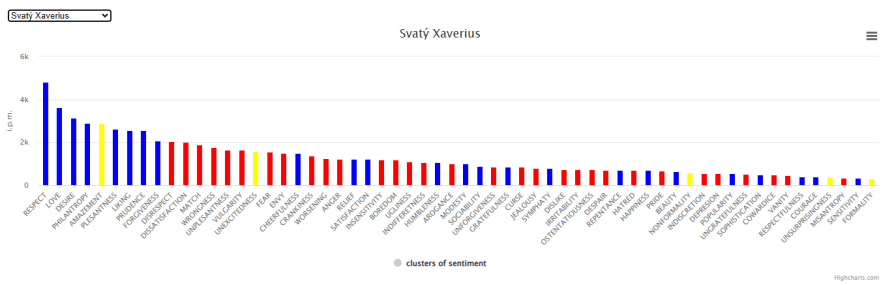


Figure 5. Sentiment analysis using a filter list created by selecting clusters from the *Thesaurus of the Czech Language*.

emotion. The starting point for the construction of the filtering database was the extensive thematic *Thesaurus of the Czech Language* (2008), which processes the Czech lexicon into 885 conceptual clusters. The clusters for emotions include the individual lemmas that were used for filtering and as the criterion of the cluster (see Fig. 5).

A newly added functionality of the project is network models of relationships between characters in the story. The active relationships between characters are determined by two basic aspects: the dialogue between characters and the narrator's relation to such dialogue, which is not explicitly realised in the text (see Fig. 6). These network models are related to the network models of the characters' movements across the fictional topography, which are part of the cartographic model section in the web application environment.

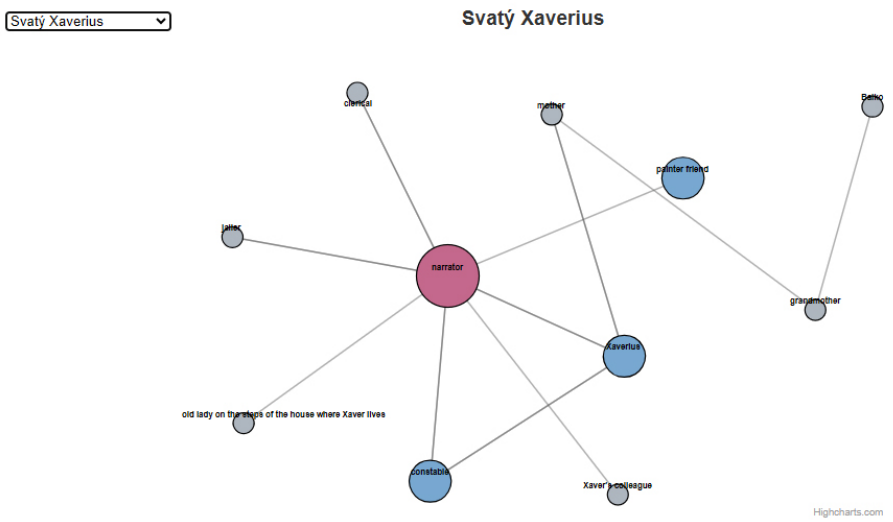


Figure 6. Relationships between the narrator and the characters in Jakub Arbes' novel *Svatý Xaverius*: the pink area is related to the narrator, the blue area is related to the character with whom the narrator or another character is speaking, the grey area is related to a character whose communication is not explicitly realised in the narrative, e.g. referred to by the narrator or another character.

Tools for basic corpus mining in terms of selected linguistic-corpus criteria are an integral part of the "prose corpus". Specifically, this includes functions for collocation search (five positions on the left and five positions on the right), word frequency and word type search, listing of specific word categories or bigram collocations, and word search in close context. This feature is based on the ability to search for two specific lemmas. If these occur in a context that is no more than two positions to the left or right, their context is listed (see Fig. 7).

These functions may serve not only linguists but also literary scholars, who can use them to query or verify other information in the corpus. For example, the word clouds model that is generated from autosemantic words and which is built for each text (see Fig. 8) can be seen as a quantitative representation of the motivic field of the text. The words in the cluster model are only distinguished by their letter size, which depends on their frequency, but to find the context, one needs to look for their concordances or collocations. This is what the other tools of the Czech prose corpus mentioned here are used for.

As a result, we aim to create a multifunctional digital web platform that would be able to offer support for individual research to both literary scholars and linguists who focus on Czech narrative literature. However, an integral part of any such research is the interpretation of the data, or, rather, the models that represent the data. We will address this issue in general terms in the final section; specific examples of the interpretation of cartographic and quantitative models are presented in the study *Literary-Cartographic Models in the Corpus of Czech Prose*, which is presented elsewhere in this issue.

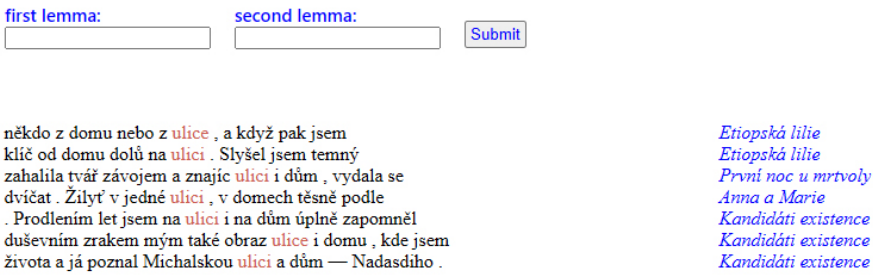


Figure 7. Words in context in Jakub Arbes' novels.



Figure 8. Word cloud for the novel *Svatý Xaverius* by Jakub Arbes.

Conclusion

As in any research field in which quantitative, statistical or cartographic models are used, so in literary science operating with these tools and modes of representation, it is true that their mere existence is not sufficient for fully fledged research, or for its final concretisation. As follows from the structure of the corpus presented here and from the possibilities of its mining, research can be carried out both at the level of specific authors, of course within the framework of cross-data comparisons, which in the future (once the corpus is data-representative) will make it possible to observe the changes between authors with respect to different areas and layers of narrative art text, and within the whole corpus. This raises the important question of how to interpret the data and the models based on it.

In terms of how to read individual models, it is appropriate at this point to refer to some of Doležel's works, which are focused on the application of "exact" methods in literary science, because, as we believe, his conclusions are timeless, valid earlier and today. In his corroboration of Alvaro Ellegård's study *Genre Styles, Individual Styles, and Authorship Identification* (Ellegård, 1982), which Doležel delivered at the 51st Nobel Symposium in Lidingö (Doležel, 1982), he states: "I cannot do more than to outline briefly the problems connected with two such strategies (both of which are generally known in quantitative investigations), namely SCALING and INTERPRETATION. The first procedure can be characterised as transformation of qualitative properties into quantitative data, while the second one is conversion of numerical data into structural description. These two strategies are indispensable for any empirical theory; the neglect of their foundations in the text study is, in my opinion, a major cause of our difficulties, misconceptions and misunderstanding. ... There are, in principle, two possibilities of interpreting numerical data: in terms of qualitative properties and in terms of quantitative structures. ... In a qualitative interpretation, the data are taken as INDICATORS (indices, symptoms), i.e. their values (or difference in values) are interpreted as signalling the presence of certain qualitative (formal) properties, relations or taxonomies" (Doležel, 1982, pp. 540–541, 543). The indices or symptoms Doležel refers to are markers signalling the structural situation of the phenomenon in question. Thus, one cannot be satisfied with mere descriptions of models in interpretation, but these should be understood as macrostructural "maps" showing the behaviour of the phenomenon being modelled. And it is this behaviour, i.e. the symptomatic or asymptomatic

nature of the phenomenon in the context of the material being analysed, that is the first step towards thinking about its nature and behaviour in a complex structural system.

It is also true that when interpreting, it is absolutely necessary to compare individual models with each other, not only models of the same type, e.g. not just cartographic or quantitative models, but also those that belong to a different model type. Each of them views the literary text from a different perspective by modelling different structural layers and relationships. It is then absolutely fundamental that any interpretation based on mathematical, cartographic or network modelling must be based on the theoretical basis of the discipline within which these models are applied. In other words, the use of such models in literary scholarship must necessarily lead primarily to a literary-scientific interpretation, which does not, of course, preclude any other potential use of such models.

Translated from Czech by
Josef Línek

References

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2012). *Digital_humanities*. MIT Press. https://doi.org/10.7551/mitpress/9248.001.0001.

Burdicková, A., Druckerová, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2019). *Digital_humanities* (D. Vichnar, Trans.). Academia.

Červenka, M., & Sgallová, K. (1995). *Z večerní školy versologie III*. Ústav pro českou literaturu Akademie věd České republiky.

Debnár, M., & Gogora, A. (2019). *Digitálne trendy v súčasných humanitných vedách*. Univerzita Konštantina Filozofa v Nitre, Centrum digitalnych humanitnych vied pri Filozofickej fakulte.

Doležel, L. (1965). Pražská škola a statistická teorie básnického jazyka. *Česká literatura*, *13*(2), 101–113.

Doležel, L. (1982). Discussion of Alvar Ellegård's paper, *Genre styles, individual styles, and autorship identification*. In S. Allén (Ed.), *Text processing: Text analysis and generation, text typology and atribution: Proceedings of Nobel Symposium 51* (pp. 539–551). Almqvist & Wiksell International.

Ellegård, A. (1982). Genre styles, individual styles, and authorship identification. In S. Allén (Ed.), *Text processing: Text analysis and generation, text typology and atribution: Proceedings of Nobel Symposium 51* (pp. 519–537). Almqvist & Wiksell International.

Jakobson, R. (1995a). K popisu Máchova verše. In R. Jakobson, *Poetická funkce* (pp. 427–476). H&H.

Jakobson, R. (1995b). Verš staročeský. In R. Jakobson, *Poetická funkce* (pp. 275–318). H&H.

Kvantitativní a statistické metody [Special issue]. (2022). *Česká literatura*, *70*(6). https://asjournals.lib.cas.cz/ceskaliteratura/archiv;pid=uid:74b544c8–9bed-4166–8bb7-c4a1a9893bd5

Levý, J. (1971). *Bude literární věda exaktní vědou?* Československý spisovatel.

Moretti, F. (2015). *Grafy, mapy, stromy: Abstraktní modely literární historie*. Karolinum.

Moretti, F. (2019). Teorie sítí a analýzy syžetu. *ArteActa*, *2019*(3), 57–84. https://doi.org/10.62804/aa.2019.003

Mukařovský, J. (1982a). Obecné zásady a vývoj novočeského verše. In J. Mukařovský, *Studie z poetiky* (pp. 365–438). Odeon.

Mukařovský, J. (1982b). Polákova vznešenost přírody. In J. Mukařovský, *Studie z poetiky* (pp. 439–513). Odeon.

Petrů, E. (1968). *Exaktní metody v literárněvědné práci*. Filosofická fakulta.

Piorecký, K. (2016). *Česká literatura a nová média*. Academia.

Roy, D., & Menon, N. (2022). No "making," not now: Decolonizing digital humanities in South Asia. In D. Fiormonte, S. Chaudhuri, & P. Ricaurte (Eds.), *Global debates in the digital humanities* (pp. 186–201). University of Minnesota Press.

Vašák, P. (1980). *Metody určování autorství*. Academia.

Změlík, R. (2015). *Kvantitativně-korpusová analýza a literární věda. Model a realizace autorského korpusu a slovníku Jana Čepa v kontextu zahraniční a české autorské lexikografie.* Vydavatelství Univerzity Palackého.

Změlík, R. (2019). *Konceptualizace barev v narativní fikci na pozadí kvantitativních modelů*. Univerzita Palackého v Olomouci.

Digital resources

Česká literární bibliografie. (n.d.). Ústav pro českou literaturu AV ČR. https://clb.ucl.cas.cz/

Český národní korpus. (n.d.). https://www.korpus.cz/

Digital humanities. (n.d.). Katedra obecné lingvistiky. https://kol.upol.cz/studium/moduly/digital-humanities/

DL4DH. (n.d.). Národní knihovna ČR. https://www.lib.cas.cz/projekt/dl4dh/

Manuscriptorium. (n.d.). https://www.manuscriptorium.com/cs

MorphoDiTa. (n.d.). https://lindat.mff.cuni.cz/services/morphodita/

Piorecký, K., & Škrabal, M. (n.d.). *Korpus současné poezie*. https://versologie.cz/ksp/web_content/

Plecháč, P. (n.d.). *Korpus českého verse*. https://versologie.cz/v2/web_content/corpus.php

Sentiment Analysis in Czech. (n.d.). https://ufal.mff.cuni.cz/sentiment/data

Stanovisko České asociace pro digitální humanitní vědy k hodnocení a uznávání výzkumných výsledků. (2023, May 15). Česká asociace pro digitální humanitní vědy. https://www.czadh.cz/stanovisko-k-hodnoceni-a-uznavani-vyzkumnych-vysledku/

Ústav pro českou literaturu AV ČR. (2020, August 12). *Digital humanities a literární věda (21. 11. 2017)* [Video]. YouTube. https://www.youtube.com/watch?v=UGPGIiCphUs

Změlík, R. (n.d.). *Literary cartographic and quantitative models of Czech novels from the 19th to 21st century*. https://www.korpusprozy.com/

Computational Literary Studies v kontekstu současné české literární vědy – zvláště konceptu korpusu české prózy

V tomto příspěvku se zaměřuji na některé otázky, které souvisí s oborem zvaným Computational Literary Studies a který náleží do oblasti Digital Humanities. V úvodu se zabývám především československou literární a částečně i lingvistickou tradicí, která této disciplíně předcházela. Poté jsou představeny české projekty, konkrétně specializované literární korpusy. Posléze je pozornost vznikajícímu korpusu a webové aplikaci korpus české prózy. Je popsána jeho celková struktura i jednotlivé jeho parametry a konkrétní funkce.

Klíčová slova: digitální literární studia, kvantitativní a kartografické modely, strukturalismus, česká literární věda, korpusová lingvistika, korpus v literární vědě.

Computational Literary Studies w kontekście współczesnych czeskich studiów literaturoznawczych – zwłaszcza koncepcji korpusu czeskiej prozy

W niniejszym artykule rozważam wybrane zagadnienia związane z będącymi częścią humanistyki cyfrowej Computational Literary Studies. We wstępie przedstawiam przede wszystkim czechosłowacką

tradycję literaturoznawczą i częściowo językoznawczą, które poprzedzały Computational Literary Studies. Następnie prezentuję czeskie projekty, w szczególności specjalistyczne korpusy literackie. Zwracam też uwagę na powstający korpus (wraz z aplikacją internetową) czeskiej prozy. Opisuję jego ogólną strukturę, poszczególne właściwości i specyficzne funkcje.

Słowa kluczowe: literaturoznawstwo cyfrowe, modele ilościowe i kartograficzne, strukturalizm, czeskie literaturoznawstwo, językoznawstwo korpusowe, korpusy w literaturoznawstwie.

Przekład z języka czeskiego
Robert Kulmiński

Note

Richard Změlík, Faculty of Art, Palacký University in Olomouc, Olomouc, Czechia.
richard.zmelik@upol.cz
https://orcid.org/0000-0002-5414-4574

The preparation and publication of this paper was made possible by financial support provided by the Faculty of Arts of Palacký University in Olomouc in the years 2022–2024 from the Fund for Support of Scientific Activities FPVC2022/20.
No competing interests have been declared.

Publication History

Received: 2023-09-27, Accepted: 2024-04-07, Published: 2024-12-23