Bohemistyka, 2025(1): 139–154 ISSN 1642–9893 ISSN (online) 2956–4425
DOI: http://doi.org/10.14746/bo.2025.1.8

On the question of the acquisition of literary texts and the method of their processing for the needs of independent literary research[1]

RICHARD ZMĚLÍK
Palacky University in Olomouc
ORCID: https://orcid.org/0000-0002-5414-4574
E-mail: richard.zmelik@upol.cz

Abstract: The study deals with the issue of acquisition of digital literary data, specifically prose texts of Czech literature, which the data would serve for independent scientific research in the context of digital humanities, or computational literary studies. In the first part, we focus on selected available foreign textual databases, which we characterize with respect to the stated goal, i.e. to the existence of such a digital data collection that would be internally structured and machine-readable. We then focus on the Czech environment, in the context of which we present the emerging database of prosaic texts of Czech literature. We describe its basic structure, the advantage of such structuring, and concrete examples of possible use of the database in statistical analysis of literary texts. We conclude that in the context of the current development of DH we can expect an increasing demand not only for specialized web applications of digital literary corpora, but especially for access to such or similar databases, as these allow for highly variable and individual research.

[1] This publication was created with the support of FPVC2022/20.

Keywords: digital literary studies, computational literary studies, literary corpora, literary databases

In the context of the current progressive development of digital humanities, a discipline that is generally focused on the functional interconnection of digital technologies and the humanities or social sciences, one of the important issues concerning the availability of reliable machine-readable structured data emerges, especially in the field of literary studies, where these approaches are still strongly reflected and applied mainly in the foreign research context. The basis of such data is formed primarily by digitized literary texts, which would form a solid basis not only for partial analyses carried out by means of machine processing, but also for larger and more representative (from a literary-historical point of view) digital literary corpora, which, in addition to the nowadays standard tools used by corpus linguistics (e.g. etc.) would be able to offer such specialized tools that would be meaningfully usable primarily for literary research, and secondarily for linguistic or other research, e.g. historical research, etc.

From the few examples we have available today we know that the basic condition for creating a set of meaningful and useful tools for mining a literary corpus is primarily a question of a functional connection between the literary science task and the real programming output.[2] However, what precedes this cooperation, or rather what it necessarily relies on, is the relevant digital data in the form of digitized literary texts, preferably in the form of structured and machine-readable fi-

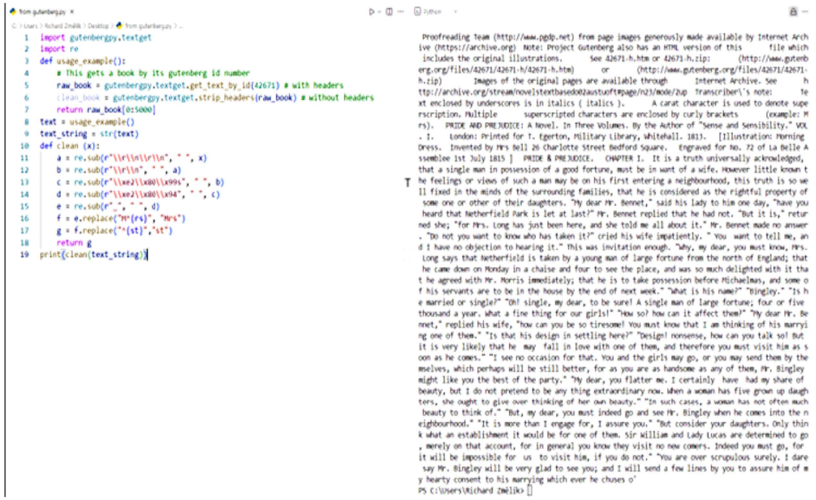
[2] Although I believe that the above stated sequence is self-evident and should not be understood in reverse order, we present this information here deliberately because where DH methods are not yet fully developed in a literary-scientific context, which is related to the critical approach to DH, questions may arise over the possibilities of adequate implementation of literary-scientific requirements in a programming context. In a scientific environment, the requirements for specific applications and software should always be primarily based on the methodological and theoretical requirements of the discipline, i.e., in this case, the literary science context. Necessary constraints on the formulation of certain requirements arise

les. Let us look at a few selected examples offered by the foreign and Czech environment. Currently, the most accessible digital database of literary texts is *Project Gutenberg*, which offers wide access to English-language literary texts (https://archive.org/details/gutenberg). These can be retrieved for machine processing in an open application environment, from where texts can be either simply manually downloaded or copied, or whole text files can be retrieved by web scraping[3].

```
   The Project Gutenberg eBook, Pride and Prejudice, by
Jane Austen, Edited by R. W. (Robert William) Chapman
This eBook is for the use of anyone anywhere at no cost and
with almost no restrictions whatsoever.  You may copy it,
give it away or re-use it under the terms of the Project
Gutenberg License included with this eBook or online at
www.gutenberg.org
Title: Pride and Prejudice
Author: Jane Austen
Editor: R. W. (Robert William) Chapman
Release Date: May 9, 2013 [eBook #42671]
Language.
Character set encoding: ISO-8859-1
***START OF THE PROJECT GUTENBERG EBOOK PRIDE AND PREJUDI-
CE***
E-text prepared by Greg Weeks, Jon Hurst, Mary Meehan, and
the Online Distributed Proofreading Team (http://www.
pgdp.net) from page images generously made available by
Internet Archive (https://archive.org)
Note: Project Gutenberg also has an HTML version of this
      file which includes the original illustrations.
      See 42671-h.htm or 42671-h.zip:
      (hHttp://www.gutenberg.org/files/42671/42671-h/42671-h.htm)
      or
      (http://www.gutenberg.org/files/42671/42671-h.zip)
      Images of the original pages are available through
      Internet Archive. See
http://archive.org/stream/novelstextbasedo02austuoft#page/n23/m
ode/2up
```

Fig. 1: Sample OCR text of Jane Austen: *Pride and Prejudice* in the Gutenberg Library (https://ia903107.us.archive.org/8/items/prideandprejudic42671gut/42671-8.txt)

Another advantage of this digital library project is the availability of the database in Python as an installable library (see https://pypi.org/project/gutenbergpy). Once the library is loaded, it is possible to work with the book titles immediately. The following example uses a simple code to demonstrate the loading of the text of Jane Austen's *Pride and Prejudice* in Python, or the display of the first 5000 characters from the OCR format.


from gutenberg.py
C: > Users > Richard Změlík > Desktop > from gutenberg.py > ...
Python

```
1  import gutenbergpy.textget
2  import re
3  def usage_example():
4      # This gets a book by its gutenberg id number
5      raw_book = gutenbergpy.textget.get_text_by_id(42671) # with headers
6      clean_book = gutenbergpy.textget.strip_headers(raw_book) # without headers
7      return raw_book[0:5000]
8  text = usage_example()
9  text_string = str(text)
10 def clean (x):
11     a = re.sub(r"\\r\\n\\r\\n", " ", x)
12     b = re.sub(r"\\r\\n", " ", a)
13     c = re.sub(r"\\xe2\\x80\\x99s", " ", b)
14     d = re.sub(r"\\xe2\\x80\\x94", " ", c)
15     e = re.sub(r"_", " ", d)
16     f = e.replace("M^{rs}", "Mrs")
17     g = f.replace("^{st}","st")
18     return g
19 print(clean(text_string))
```

```
 Proofreading Team (http://www.pgdp.net) from page images generously made available by Internet Arch
ive (https://archive.org)  Note: Project Gutenberg also has an HTML version of this       file which
 includes the original illustrations.       See 42671-h.htm or 42671-h.zip:       (http://www.gutenb
erg.org/files/42671/42671-h/42671-h.htm)       or       (http://www.gutenberg.org/files/42671/42671-
h.zip)       Images of the original pages are available through       Internet Archive. See       h
ttp://archive.org/stream/novelstextbasedo02austuoft#page/n23/mode/2up  Transcriber\'s note:       Te
xt enclosed by underscores is in italics ( italics ).       A carat character is used to denote supe
rscription. Multiple       superscripted characters are enclosed by curly brackets       (example: M
rs).   PRIDE AND PREJUDICE: A Novel. In Three Volumes. By the Author of "Sense and Sensibility." VOL
. I.    London: Printed for T. Egerton, Military Library, Whitehall. 1813.   [Illustration: Morning
Dress.  Invented by Mrs Bell 26 Charlotte Street Bedford Square.   Engraved for No. 72 of La Belle A
ssemblee 1st July 1815 ]   PRIDE & PREJUDICE.   CHAPTER I.  It is a truth universally acknowledged,
that a single man in possession of a good fortune, must be in want of a wife. However little known t
he feelings or views of such a man may be on his first entering a neighbourhood, this truth is so we
ll fixed in the minds of the surrounding families, that he is considered as the rightful property of
 some one or other of their daughters. "My dear Mr. Bennet," said his lady to him one day, "have you
 heard that Netherfield Park is let at last?" Mr. Bennet replied that he had not. "But it is," retur
ned she; "for Mrs. Long has just been here, and she told me all about it." Mr. Bennet made no answer
. "Do not you want to know who has taken it?" cried his wife impatiently. " You  want to tell me, an
d I have no objection to hearing it." This was invitation enough. "Why, my dear, you must know, Mrs.
 Long says that Netherfield is taken by a young man of large fortune from the north of England; that
 he came down on Monday in a chaise and four to see the place, and was so much delighted with it tha
t he agreed with Mr. Morris immediately; that he is to take possession before Michaelmas, and some o
f his servants are to be in the house by the end of next week." "What is his name?" "Bingley." "Is h
e married or single?" "Oh! single, my dear, to be sure! A single man of large fortune; four or five
thousand a year. What a fine thing for our girls!" "How so? how can it affect them?" "My dear Mr. Be
nnet," replied his wife, "how can you be so tiresome! You must know that I am thinking of his marryi
ng one of them." "Is that his design in settling here?" "Design! nonsense, how can you talk so! But
it is very likely that he  may  fall in love with one of them, and therefore you must visit him as s
oon as he comes." "I see no occasion for that. You and the girls may go, or you may send them by the
mselves, which perhaps will be still better, for as you are as handsome as any of them, Mr. Bingley
might like you the best of the party." "My dear, you flatter me. I certainly  have  had my share of
beauty, but I do not pretend to be any thing extraordinary now. When a woman has five grown up daugh
ters, she ought to give over thinking of her own beauty." "In such cases, a woman has not often much
 beauty to think of." "But, my dear, you must indeed go and see Mr. Bingley when he comes into the n
eighbourhood." "It is more than I engage for, I assure you." "But consider your daughters. Only thin
k what an establishment it would be for one of them. Sir William and Lady Lucas are determined to go
, merely on that account, for in general you know they visit no new comers. Indeed you must go, for
it will be impossible for  us  to visit him, if you do not." "You are over scrupulous surely. I dare
 say Mr. Bingley will be very glad to see you; and I will send a few lines by you to assure him of m
y hearty consent to his marrying which ever he chuses o'
PS C:\Users\Richard Změlík> 
```

Fig. 2: Strings that are irrelevant to the text analysis are deleted from the text (see lines 11–17) in the left panel. The retrieval of a specific text is done via an ID that is identical to the eBook number in the Gutenberg database (cf. line 10 in Fig. 1 and line 5 here in the left panel)

naturally in the context of humanities. For example, until recently it was unthinkable to require a machine algorithm to interpret a continuous literary text, but this is beginning to change with the advent of AI. The next question, of course, is to what extent such an interpretation is currently professionally valid.

[3] These are methods of downloading web content using a specially written program that searches for content by structural html tags.

The modified text can already be treated as a data type (here string), which is important for further machine analysis. Although the Gutenberg project, as well as its Python module, is a unique project that allows to analyse a large number of English literary titles and thus to perform a representative distant reading analysis, which has a more comprehensive predictive value reflecting the situation in English written prose in a longer period of time, but as can be seen from the example, this is not structured data. The metadata are thus directly part of the text, and it is therefore necessary to filter them out for further work (author's name, place and year of publication, publisher's notes, etc.). How can such a text be further processed? Specifically, it can be an analysis of lexicon, word richness, motives, themes, but also a stylometric analysis showing the statistical distances between the texts under study and, of course, many other criteria. Some of them will be briefly demonstrated here.

Another current major project for a dedicated digital database is the *DraCor* project at the University of Potsdam. It focuses on the dramatic texts of selected German (485 texts) and Russian (80 texts) plays written between 1730 and 1930. The specificity of this project lies in the fact that it is not a simple database of digitized texts, but a much more sophisticated digital environment that uses the so-called "corpus" to work with corpora. API approaches that operate on data provided in the form of the Json format, which is, among other things, suitable for data and metadata storage and structuring (see below). As the authors state,

> DraCor is not primarily a website. DraCor is a showcase for the concept of Programmable Corpora. [...] DraCor is an ecosystem. You can connect to it on different levels. The corpora can be used freely, even independently of the platform itself. For ease of use, there is an interface providing access to specific slices of the corpora, i.e. research data that can be used directly in your own work [...] DraCor aims to create an interface between traditional and digital literary studies. The extent to which you can involve DraCor in your research on European drama (or literature in general), depends on your level of technical expertise (or support) (https://dracor.org/doc/what-is-dracor#).

The information given in the project description is of interest primarily because it demonstrates how today's digital corpus-database environments are being built, and in particular what their nature is. In other words, users are not reliant on installed corpus search tools, as we are used to with the "traditional" corpora that have been developed since the 1990s, but can design (i.e. program) them themselves according to their own research needs. Therefore, the authors draw attention to the necessary technical (i.e. programming) skills or support. This is an indisputable advantage of a corpus-database conceived in this way, including the possibility to further extend it. This current trend can also be observed in other digital corpora; if we limit ourselves to the Czech environment, we can mention the *Czech National Corpus* (https://www.korpus.cz) or the *Czech Verse Corpus* (https://versologie.cz/v2/web_content/corpus.php).

This approach is also progressive compared to other foreign projects. A list of other digital literary corpora can be found in the CLARIN research infrastructure database, which serves as a platform for linguistic, social and cultural data (https://www.clarin.eu/resource-families/ literary-corpora). For the sake of comparison, we select a few examples offered here. For example, *The Complete Corpus of Anglo-Saxon Poetry* (https://sacred-texts.com/neu/ascp) offers a simple, albeit extensive, digitized database of Anglo-Saxon verse texts, similarly the *Collection of older original Estonian-language works of fiction* is a collection of digitized literary texts of Estonian provenance. The project also has the possibility of browsing through older editions, a timeline of historical events that form the context of Estonian works. The advantage of the project, as its authors state, is that it currently includes all 19th century Estonian artistic literature, which is important not only in terms of preserving historical texts and documents, but can also serve as a complete textual resource for systematic research. Other corpora, such as *Classics of English and American Literature in Finnish (CEAL)* (https://korp.csc.fi/shibboleth-ds/index.html?https%3A%2F%2Fkorp.csc.fi%2Fkorp%2F%3Fsaved_params%3D1705922341337) or *Classics of Finnish Literature, Kielipankki Version* (https://korp.csc.fi/korp/#?prequery_within=sentence&cqp

=%5B%5D&corpus=skk_aho,skk_canth,skk_finne,skk_jarnefelt,skk_kailas,skk_lassila,skk_linnankoski,skk_kramsu,skk_lehtonen,skk_leino,skk_pakkala,skk_siljo,skk_sodergran,skk_wilkuna), contain some eselected tools for searching the corpus, which is made possible by processing the textual data at the level of syntactic parsing and morphological tagging. If we look at other corpora, then majority of them are either implemented as digital collections of literary texts, but also of other texts, such as essays, or the texts are processed at the level of parsing and morphological annotation. It is therefore probably important to distinguish between a digital collection, which is the first example, and corpora that already have certain corpus tools for searching.

There are numerous of similar projects today. We would like to mention one of them here, because it is also related to the Czech environment. It is the *European Literary Text Collection (ELTeC)* project (https://distantreading.github.io/ELTeC), which, as the name implies, focuses on digitized collections of literary works of European literatures, which it offers in the form of web content. The condition for the creation of a specific digital collection is that it must contain exactly 100 book titles. Focusing on the Czech collection (https://distantreading.github.io/ELTeC/cze/index.html), the first thing that is somewhat surprising is the selection of the texts themselves. The first problem for a possible systematic analysis is the non-representativeness of the database. The texts that are part of it are mostly not part of the Czech literary canon; in many cases they are marginal and less known or unknown 19th century texts, in some cases, on the contrary, they are texts representing the literary canon of the 19th century Czech literature. However, the question of canonicity might not be so binding in the final analysis if there were enough texts representing the relevant historical developmental stages of Czech literature. We have in mind (sub)sets of texts relating, for example, to the 1830s, 1840s, 1850s and to other decades of the 19th century. Instead, these are mostly randomly selected texts from the period 1855–1920. Another problem is that each author is represented here by a single specific text, while for example Alois Jirásek has three texts. Such a selection de facto prevents more meaningful comparisons, the results of which would have

some more valid cognitive value for modelling the literary-historical processes of Czech literature.

Other resources that concentrate a disproportionately larger number of Czech literary works include the Kramerius database managed by the National Library of the Czech Republic (https://kramerius.Nkp.cz/kramerius/welcome.do;jsessionid=EDA92C4CCCFBB6ED5585E62C91C37BD3). This database can be used to search a truly large number of literary documents, but even using this environment has its pitfalls. Texts are collected here in pdf format, where it is always possible to download only the maximum batch of 20 pages. As a result, it is then necessary to merge the individual pdf files into a single file, perform OCR and then check, clean and save the text in a format that can be further processed by machine. The newer version of this database, Kramerius5, does offer a text transcription of the displayed pdf, i.e. a specific page from the document, but again it is necessary to manually copy the OCR part and de facto assemble the whole work in pieces, which is a very tedious and inefficient work. Not even the digital library of the Moravian Library offers a fundamentally different approach. It should be noted, however, that in the case of both these institutions, they are limited by copyright and are not primarily oriented towards the creation of databases for scientific purposes, i.e. datasets that could be further processed in the context of digital humanities research. However, we mention both resources here because their collection constitutes the largest database of Czech literary texts in the Czech Republic.

The situation is different in the case of the *Corpus of Czech Verse*, which is implemented at the Institute for Czech Literature of the Academy of Sciences of the Czech Republic. In addition to the web interface (https://versologie.cz/v2/web_content/tools.php?lang=cz), where it is possible to search according to selected versological criteria, it also offers open data (https://github.com/versotym/corpusCzechVerse) for further individual processing.

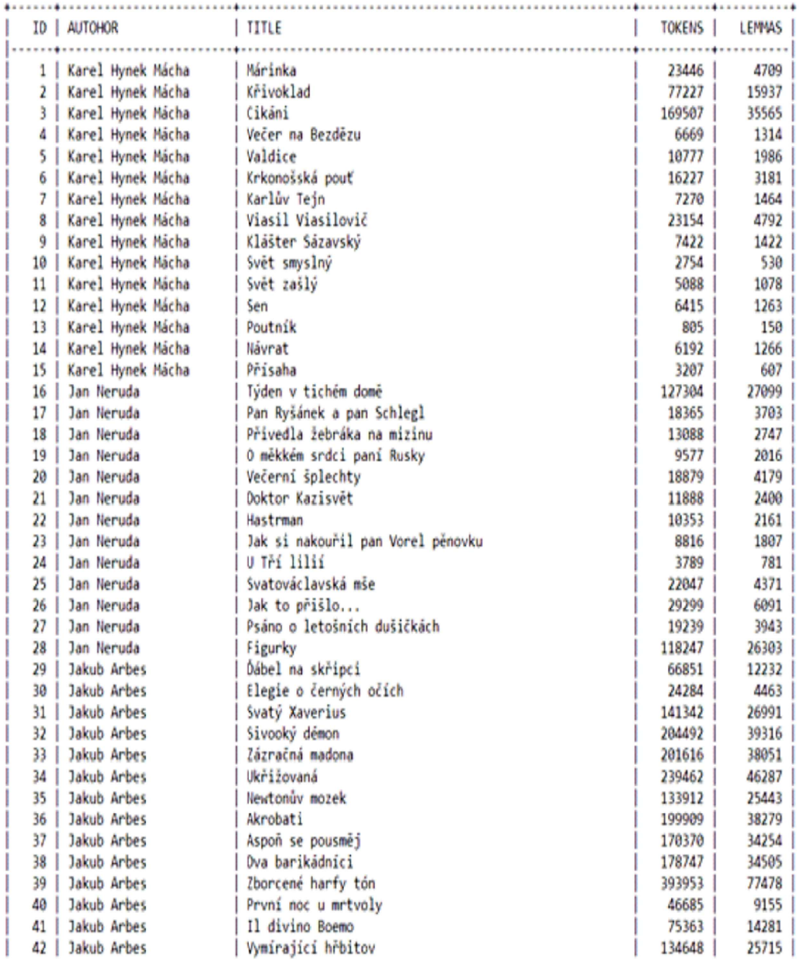
We are following a similar path in the *Literary Cartographic and Quantitative Models of Czech Novels from the 19th to 21st Century* project (https://korpusprozy.com), providing a structured and machine-readable dataset of texts and other metadata for independent research as part of the globally shared open data trend. In doing so, we aim not only to meet the generally shared call for open and accessible data, but also – as the above mentioned review of Czech text databases has shown – to provide researchers with a free and structured database of literary texts for their professional work, which can become the basis for individually focused and independent research work.

The default format in which text data is stored and provided is the Json text format, in which both text and metadata are structured into a dictionary data type (see Fig. 3). The basic principle of this data type is that it contains two values written in the manner {key: value}. If we look at the processing structure of each literary text (see Fig. 4) we can see that a dictionary can contain another dictionary, etc., which makes this data type very suitable for structuring and hierarchizing. In this case, a value is associated with the TITLE key, which is a dictionary that contains a series of keys and values.

This formatted data can be worked with completely independently, taking into account the wide variability of possible research tasks. Here we demonstrate several such examples that illustrate the different possibilities of processing such structured data. The following list (see Fig. 4) is a basic listing of works, authors, number of tokens and lemmas in each text in a summary table.

```
{
        "TITLE"     : {
                              "AUTHOR" : author name,
                              "BORN" : date of author born,
                              "DEATH" : date of author death,
                              "1. PUB PUB" : 1. publiciton of the title,
                              "ACTUAL PUB" : actual publication,
                              "TEXT" : text of title (tokens)
                              "LEMMA" : lemmas,
                              "MORPHO TAGS" : morphological token tags
                      }
}
```

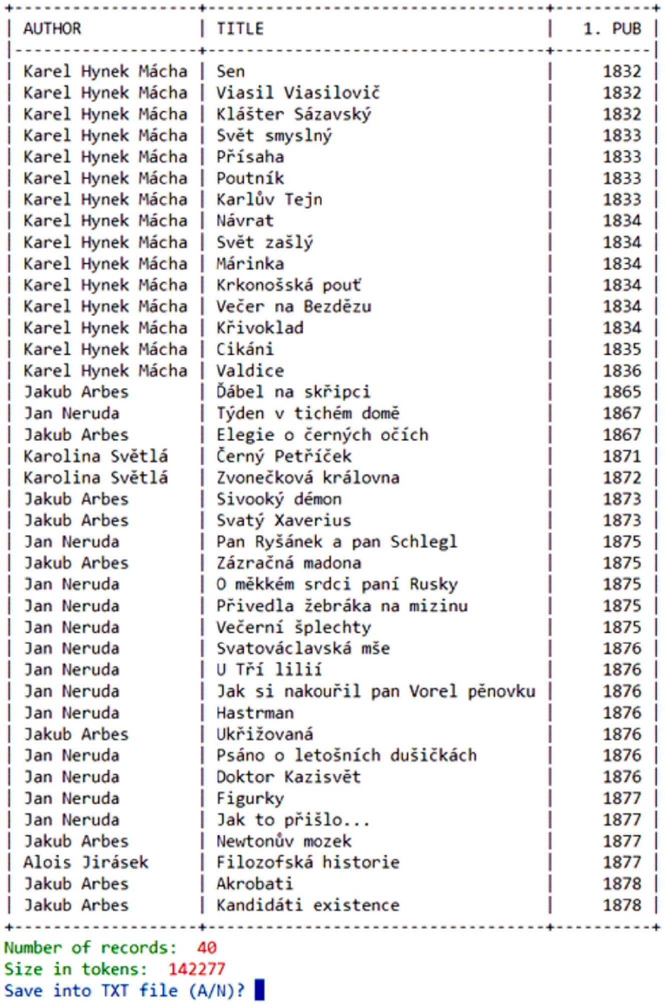
Fig. 3: Processing structure of each work



```
+------+------------------------+------------------------------------------+---------+--------+
|   ID | AUTOHOR                | TITLE                                    |  TOKENS | LEMMAS |
|------+------------------------+------------------------------------------+---------+--------|
|    1 | Karel Hynek Mácha      | Márinka                                  |   23446 |   4709 |
|    2 | Karel Hynek Mácha      | Křivoklad                                |   77227 |  15937 |
|    3 | Karel Hynek Mácha      | Cikáni                                   |  169507 |  35565 |
|    4 | Karel Hynek Mácha      | Večer na Bezdězu                         |    6669 |   1314 |
|    5 | Karel Hynek Mácha      | Valdice                                  |   10777 |   1986 |
|    6 | Karel Hynek Mácha      | Krkonošská pouť                          |   16227 |   3181 |
|    7 | Karel Hynek Mácha      | Karlův Tejn                              |    7270 |   1464 |
|    8 | Karel Hynek Mácha      | Viasil Viasilovič                        |   23154 |   4792 |
|    9 | Karel Hynek Mácha      | Klášter Sázavský                         |    7422 |   1422 |
|   10 | Karel Hynek Mácha      | Svět smyslný                             |    2754 |    530 |
|   11 | Karel Hynek Mácha      | Svět zašlý                               |    5088 |   1078 |
|   12 | Karel Hynek Mácha      | Sen                                      |    6415 |   1263 |
|   13 | Karel Hynek Mácha      | Poutník                                  |     805 |    150 |
|   14 | Karel Hynek Mácha      | Návrat                                   |    6192 |   1266 |
|   15 | Karel Hynek Mácha      | Přísaha                                  |    3207 |    607 |
|   16 | Jan Neruda             | Týden v tichém domě                      |  127304 |  27099 |
|   17 | Jan Neruda             | Pan Ryšánek a pan Schlegl                |   18365 |   3703 |
|   18 | Jan Neruda             | Přivedla žebráka na mizinu               |   13088 |   2747 |
|   19 | Jan Neruda             | O měkkém srdci paní Rusky                |    9577 |   2016 |
|   20 | Jan Neruda             | Večerní šplechty                         |   18879 |   4179 |
|   21 | Jan Neruda             | Doktor Kazisvět                          |   11888 |   2400 |
|   22 | Jan Neruda             | Hastrman                                 |   10353 |   2161 |
|   23 | Jan Neruda             | Jak si nakouřil pan Vorel pěnovku        |    8816 |   1807 |
|   24 | Jan Neruda             | U Tří lilií                              |    3789 |    781 |
|   25 | Jan Neruda             | Svatováclavská mše                       |   22047 |   4371 |
|   26 | Jan Neruda             | Jak to přišlo...                         |   29299 |   6091 |
|   27 | Jan Neruda             | Psáno o letošních dušičkách              |   19239 |   3943 |
|   28 | Jan Neruda             | Figurky                                  |  118247 |  26303 |
|   29 | Jakub Arbes            | Ďábel na skřipci                         |   66851 |  12232 |
|   30 | Jakub Arbes            | Elegie o černých očích                   |   24284 |   4463 |
|   31 | Jakub Arbes            | Svatý Xaverius                           |  141342 |  26991 |
|   32 | Jakub Arbes            | Sivooký démon                            |  204492 |  39316 |
|   33 | Jakub Arbes            | Zázračná madona                          |  201616 |  38051 |
|   34 | Jakub Arbes            | Ukřižovaná                               |  239462 |  46287 |
|   35 | Jakub Arbes            | Newtonův mozek                           |  133912 |  25443 |
|   36 | Jakub Arbes            | Akrobati                                 |  199909 |  38279 |
|   37 | Jakub Arbes            | Aspoň se pousměj                         |  170370 |  34254 |
|   38 | Jakub Arbes            | Dva barikádníci                          |  178747 |  34505 |
|   39 | Jakub Arbes            | Zborcené harfy tón                       |  393953 |  77478 |
|   40 | Jakub Arbes            | První noc u mrtvoly                      |   46685 |   9155 |
|   41 | Jakub Arbes            | Il divino Boemo                          |   75363 |  14281 |
|   42 | Jakub Arbes            | Vymírající hřbitov                       |  134648 |  25715 |
```

Fig. 4: List of the first 42 works of Czech prose writers of the 19th century with the size of each text in number of tokens and lemmas. The database currently contains 74 titles of Czech prose from the 19th century to the 21st century[20].

Another possibility resulting from the dataset is a limited listing of works that, for example, meet a certain condition. This is the time limitation for texts that were first published between 1830 and 1880 (see Fig. 5).

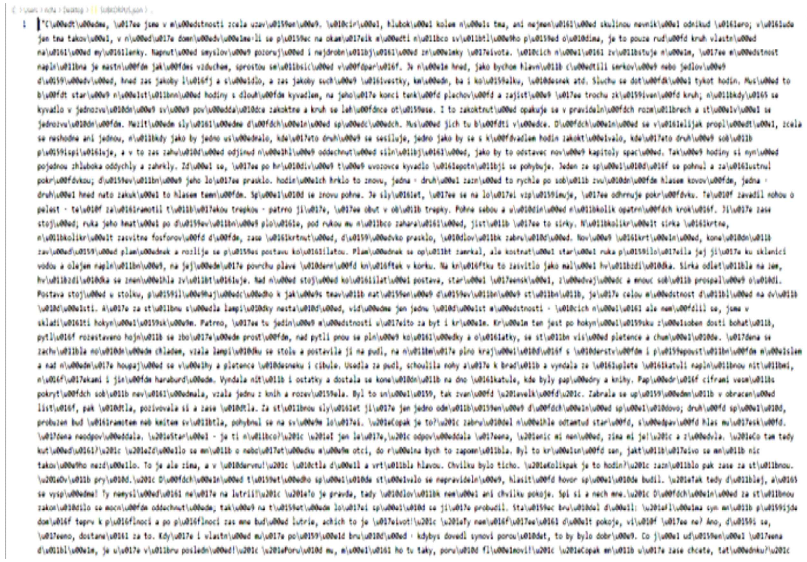


```
+--------------------+------------------------------------+---------+
| AUTHOR             | TITLE                              |  1. PUB |
|--------------------+------------------------------------+---------|
| Karel Hynek Mácha  | Sen                                |    1832 |
| Karel Hynek Mácha  | Viasil Viasilovič                  |    1832 |
| Karel Hynek Mácha  | Klášter Sázavský                   |    1832 |
| Karel Hynek Mácha  | Svět smyslný                       |    1833 |
| Karel Hynek Mácha  | Přísaha                            |    1833 |
| Karel Hynek Mácha  | Poutník                            |    1833 |
| Karel Hynek Mácha  | Karlův Tejn                        |    1833 |
| Karel Hynek Mácha  | Návrat                             |    1834 |
| Karel Hynek Mácha  | Svět zašlý                         |    1834 |
| Karel Hynek Mácha  | Márinka                            |    1834 |
| Karel Hynek Mácha  | Krkonošská pouť                    |    1834 |
| Karel Hynek Mácha  | Večer na Bezdězu                   |    1834 |
| Karel Hynek Mácha  | Křivoklad                          |    1834 |
| Karel Hynek Mácha  | Cikáni                             |    1835 |
| Karel Hynek Mácha  | Valdice                            |    1836 |
| Jakub Arbes        | Ďábel na skřipci                   |    1865 |
| Jan Neruda         | Týden v tichém domě                |    1867 |
| Jakub Arbes        | Elegie o černých očích             |    1867 |
| Karolina Světlá    | Černý Petříček                     |    1871 |
| Karolina Světlá    | Zvonečková královna                |    1872 |
| Jakub Arbes        | Sivooký démon                      |    1873 |
| Jakub Arbes        | Svatý Xaverius                     |    1873 |
| Jan Neruda         | Pan Ryšánek a pan Schlegl          |    1875 |
| Jakub Arbes        | Zázračná madona                    |    1875 |
| Jan Neruda         | O měkkém srdci paní Rusky          |    1875 |
| Jan Neruda         | Přivedla žebráka na mizinu         |    1875 |
| Jan Neruda         | Večerní šplechty                   |    1875 |
| Jan Neruda         | Svatováclavská mše                 |    1876 |
| Jan Neruda         | U Tří lilií                        |    1876 |
| Jan Neruda         | Jak si nakouřil pan Vorel pěnovku  |    1876 |
| Jan Neruda         | Hastrman                           |    1876 |
| Jakub Arbes        | Ukřižovaná                         |    1876 |
| Jan Neruda         | Psáno o letošních dušičkách        |    1876 |
| Jan Neruda         | Doktor Kazisvět                    |    1876 |
| Jan Neruda         | Figurky                            |    1877 |
| Jan Neruda         | Jak to přišlo...                   |    1877 |
| Jakub Arbes        | Newtonův mozek                     |    1877 |
| Alois Jirásek      | Filozofská historie                |    1877 |
| Jakub Arbes        | Akrobati                           |    1878 |
| Jakub Arbes        | Kandidáti existence                |    1878 |
+--------------------+------------------------------------+---------+
Number of records:  40
Size in tokens:   142277
Save into TXT file (A/N)?
```

Fig. 5: Table of titles that match the given condition, i.e. were first published between 1830 and 1880. Below the table, the total number of texts retrieved and the size of such a corpus in token counts are given.[4]

[4] The database can be downloaded after registration here: https://korpusprozy.com.

As we can see, the output is a table of literary texts sorted by the year of the first publication of the respective texts. The particular program that we use to access the database will allow us to save this selection as a single TXT text file containing the texts of all the filtered titles; of course, it is up to each user to customize their own program. It is certainly possible to save the selection as a custom Json format, Excel spreadsheet, etc. The filtered texts can be further analysed in any way, e.g. within the framework of methods standardly used in NLP. This example also shows that each user can generate his own text files (corpora) and perform various statistical measurements between them. This generation can be varied in any way. In addition to the time range, custom sub-corpora can be defined, e.g. by author names. The following figure is an example of a simple storage of all prosaic texts by Jan Neruda that are part of the database. As can be seen, the specificity of this format is its machine readability. Currently, it is perhaps one of the most widely used formats for storing and exchanging data in the digital environment of the web.



```
"C\u00edt\u00edme, \u017ee jsme v m\u00edstnosti zcela uzav\u0159en\u00e9. \u010cir\u00e1, hlubok\u00e1 kolem n\u00e1s tma, ani nejmen\u0161\u00ed skulinou nevnik\u00e1 odnikud \u0161ero; v\u0161ude
jen tma takov\u00e1, v n\u00ed\u017e domn\u00edv\u00e1me-li se p\u0159ec na okam\u017eik m\u00edti n\u011bco sv\u011btl\u00e9ho p\u0159ed o\u010dima, je to pouze rud\u00fd kruh vlastn\u00ed
na\u0161\u00ed my\u0161lenky. Napnut\u00ed smyslov\u00e9 pozoruj\u00ed i nejdrobn\u011bj\u0161\u00ed zn\u00e1mky \u017eivota. \u010cich n\u00e1\u0161 zv\u011bstuje n\u00e1m, \u017ee m\u00edstnost
napln\u011bna je mastn\u00fdm jak\u00fdms vzduchem, sprostou sm\u011bsic\u00ed v\u00fdpar\u016f. Je n\u00e1m hned, jako bychom hlavn\u011b c\u00edtili serkov\u00e9 nebo jedlov\u00e9
d\u0159\u00edv\u00ed, hned zas jakoby l\u016fj a s\u00e1dlo, a zas jakoby such\u00e9 \u0161vestky, km\u00edn, ba i ko\u0159alku, \u010desnek atd. Sluchu se dot\u00fdk\u00e1 tykot hodin. Mus\u00ed to
b\u00fdt star\u00e9 n\u00e1st\u011bnn\u00ed hodiny s dlouh\u00fdm kyvadlem, na jeho\u017e konci tenk\u00fd plechov\u00fd a zajist\u00e9 \u017ee trochu zk\u0159iven\u00fd kruh; n\u011bkdy\u0165 se
kyvadlo v jednozvu\u010dn\u00e9 sv\u00e9 pov\u00edda\u010dce zakoktne a kruh se leh\u00fdnce ot\u0159ese. I to zakoktnut\u00ed opakuje se v pravideln\u00fdch rozm\u011brech a st\u00e1v\u00e1 se
jednozvu\u010dn\u00fdm. Mezit\u00edm sly\u0161\u00edme d\u00fdch\u00e1n\u00ed sp\u00edc\u00edch. Mus\u00ed jich tu b\u00fdti v\u00edce. D\u00fdch\u00e1n\u00ed se v\u0161elijak propl\u00edt\u00e1, zcela
se neshodne ani jednou, n\u011bkdy jako by jedno us\u00ednalo, kde\u017eto druh\u00e9 se sesiluje, jedno jako by se s k\u00fdvadlem hodin zakokt\u00e1valo, kde\u017eto druh\u00e9 sob\u011b
p\u0159ispi\u0161uje, a v to zas zahu\u010d\u00ed odjinud n\u00e1hl\u00e9 oddechnut\u00ed siln\u011bj\u0161\u00ed, jako by to odstavec nov\u00e9 kapitoly spac\u00ed. Tak\u00e9 hodiny si myn\u00ed
pojednou zhluboka oddychly a zahrkly. Zd\u00e1 se, \u017ee po hr\u010div\u00e9 t\u00e9 uvozovce kyvadlo \u0161epotn\u011bji se pohybuje. Jeden ze sp\u00e1\u010d\u016f se pohnul a za\u0161ustnul
pokr\u00fdvkou; d\u0159ev\u011bn\u00e9 jeho lo\u017ee prasklo. hodin\u00e1ch hrklo to znovu, jedna - druh\u00e1 zazn\u00ed to rychle po sob\u011b zvu\u010dn\u00fdm hlasem kovov\u00fdm, jedna -
druh\u00e1 hned nato zakuk\u00e1 to hlasem temn\u00fdm. Sp\u00e1\u010d se znovu pohne. Je sly\u0161et, \u017ee se na lo\u017ei vzp\u0159imuje, \u017ee odhrnuje pokr\u00fdvku. Te\u010f zavadil nohou o
pelest - te\u010f za\u0161ramotil t\u011b\u017ekou trepkou - patrno ji\u017e, \u017ee obut v ob\u011b trepky. Pohne sebou a u\u010din\u00ed n\u011bkolik opatrn\u00fdch krok\u016f. Ji\u017e zase
stoj\u00ed; ruka jeho hmat\u00e1 po d\u0159ev\u011bn\u00e9 plo\u0161e, pod rukou mu n\u011bco zahara\u0161\u00ed, jist\u011b \u017ee to sirky. N\u011bkolikr\u00e1t sirka \u0161krtne,
n\u011bkolikr\u00e1t zasvitne fosforov\u00fd d\u00fdm, zase \u0161krtnut\u00ed, d\u0159\u00edvko prasklo, \u010dlov\u011bk zabru\u010d\u00ed. Nov\u00e9 \u0161krt\u00e1n\u00ed, kone\u010dn\u011b
zav\u00ed\u0159\u00ed plam\u00ednek a rozlije se p\u0159es postavu ko\u0161ilatou. Plam\u00ednek se op\u011bt zamrkal, ale kostnat\u00e1 star\u00e1 ruka p\u0159ilo\u017eila jej ji\u017e ku sklenici
vodou a olejem napln\u011bn\u00e9, na jej\u00ed\u017e povrchu plave \u010dern\u00fd kn\u016ftek v korku. Na kn\u016ftku to zasvitlo jako mal\u00e1 hv\u011bzdi\u010dka. Sirka odlet\u011bla na zem,
hv\u011bzdi\u010dka se zn\u00e1hla zv\u011bt\u0161uje. Nad n\u00ed stoj\u00ed ko\u0161ilat\u00e1 postava, star\u00e1 \u017eensk\u00e1, z\u00edvaj\u00edc a mnouc sob\u011b prospal\u00e9 o\u010di.
Postava stoj\u00ed u stolku, p\u0159il\u00e9haj\u00edc\u00edho k jak\u00e9s tmav\u011b nat\u0159en\u00e9 d\u0159ev\u011bn\u00e9 st\u011bn\u011b, je\u017e celou m\u00edstnost d\u011bl\u00ed na dv\u011b
\u010d\u00e1sti. A\u017e za st\u011bnu s\u00edla lampi\u010dky nesta\u010d\u00ed, vid\u00edme jen jednu \u010d\u00e1st m\u00edstnosti - \u010cich n\u00e1\u0161 ale nem\u00fdlil se, jsme v
skladi\u0161ti hokyn\u00e1\u0159sk\u00e9m. Patrno, \u017ee tu jedin\u00e9 m\u00edstnosti u\u017e\u00edto za byt i kr\u00e1m. Kr\u00e1m ten jest po hokyn\u00e1\u0159sku z\u00e1soben dosti bohat\u011b,
pytl\u016f rozestaveno hojn\u011b se zbo\u017e\u00edm prost\u00fdm, nad pytli pnou se pln\u00e9 ko\u0161\u00edky a o\u0161atky, se st\u011bn vis\u00ed pletence a chum\u00e1\u010de. \u017dena se
zachv\u011bla no\u010dn\u00edm chladem, vzala lampi\u010dku se stolu a postavila ji na pudl, na n\u011bm\u017e plno kraj\u00e1\u010d\u016f s \u010derstv\u00fdm i p\u0159epoust\u011bn\u00fdm m\u00e1slem
a nad n\u00edm\u017e houpaj\u00ed se v\u00e1hy a pletence \u010desneku i cibule. Usedla za pudl, schoulila nohy a\u017e k brad\u011b a vyndala ze \u0161uplete \u0161katuli napln\u011bnou nit\u011bmi,
n\u016f\u017ekami i jin\u00fdm haraburd\u00edm. Vyndala nit\u011b i ostatky a dostala se kone\u010dn\u011b na dno \u0161katule, kde byly pap\u00edry a knihy. Pap\u00edr\u016f ciframi vesm\u011bs
pokryt\u00fdch sob\u011b nev\u0161\u00edmala, vzala jednu z knih a rozev\u0159ela. Byl to sn\u00e1\u0159, tak zvan\u00fd \u201evelk\u00fd\u201c. Zabrala se up\u0159\u00edmn\u011b v obracen\u00ed
list\u016f, pak \u010dtla, pozivovala si a zase \u010dtla. Za st\u011bnou sly\u0161et ji\u017e jen jedno odm\u011b\u0159en\u00e9 d\u00fdch\u00e1n\u00ed sp\u00e1\u010dovo; druh\u00fd sp\u00e1\u010d,
probuzen bud \u0161ramotem neb knitem sv\u011btla, pohybnul se na sv\u00e9m lo\u017ei. \u201eCopak je to?\u201c zabru\u010del n\u00e1hle odtamtud star\u00fd, s\u00edpav\u00fd hlas mu\u017esk\u00fd.
\u017dena neodpov\u00eddala. \u201eStar\u00e1 - je ti n\u011bco?\u201c \u201eI jen le\u017e,\u201c odpov\u00eddala \u017eena, \u201enic mi nen\u00ed, zima mi je!\u201c a z\u00edvla. \u201eCo tam tedy
kut\u00ed\u0161?\u201c \u201eZd\u00e1lo se mn\u011b o nebo\u017et\u00edku m\u00e9m otci, do r\u00e1na bych to zapomn\u011bla. Byl to kr\u00e1sn\u00fd sen, jakt\u011b\u017eivo se mn\u011b nic
takov\u00e9ho nezd\u00e1lo. To je ale zima, a v \u010dervnu!\u201c \u010ctla d\u00e1l a vrt\u011bla hlavou. Chvilku bylo ticho. \u201eKolikpak je to hodin?\u201c zazn\u011blo pak zase za st\u011bnou.
\u201eDv\u011b pry\u010d.\u201c D\u00fdch\u00e1n\u00ed t\u0159et\u00edho sp\u00e1\u010de st\u00e1valo se nepravideln\u00e9, hlasit\u00fd hovor sp\u00e1\u010de budil. \u201eTak tedy d\u011blej, a\u0165
se vysp\u00edme! Ty nemysl\u00ed\u0161 ne\u017e na lutrii!\u201c \u201eTo je pravda, tady \u010dlov\u011bk nem\u00e1 ani chvilku pokoje. Spi si a nech mne.\u201c D\u00fdch\u00e1n\u00ed za st\u011bnou
zakom\u010dilo se mocn\u00fdm oddechnut\u00edm; tak\u00e9 na t\u0159et\u00edm lo\u017ei sp\u00e1\u010d se ji\u017e probudil. Sta\u0159ec bru\u010del d\u00e1l: \u201eFl\u00e1ma syn mn\u011b p\u0159ijde
dom\u016f teprv k p\u016flnoci a po p\u016flnoci zas mne bud\u00ed lutrie, achich to je \u017eivot!\u201c \u201eTy nem\u016f\u017ee\u0161 d\u00e1t pokoje, vi\u010f \u017ee ne? Ano, d\u0159\u00ed se,
\u017eeno, dostane\u0161 za to. Kdy\u017e i vlastn\u00ed mu\u017e po\u0159\u00e1d bru\u010d\u00ed - kdybys dovedl synovi porou\u010det, to by bylo dobr\u00e9. Co j\u00e1 ud\u0159en\u00e1 \u017eena
d\u011bl\u00e1m, je u\u017e v\u011bru posledn\u00ed!\u201c \u201ePoru\u010d mu, m\u00e1\u0161 ho tu taky, poru\u010d fl\u00e1movi!\u201c \u201eCopak mn\u011b u\u017e zase chcete, tat\u00ednku?\u201c
```

Fig. 6: Example of saving text data into Json format.

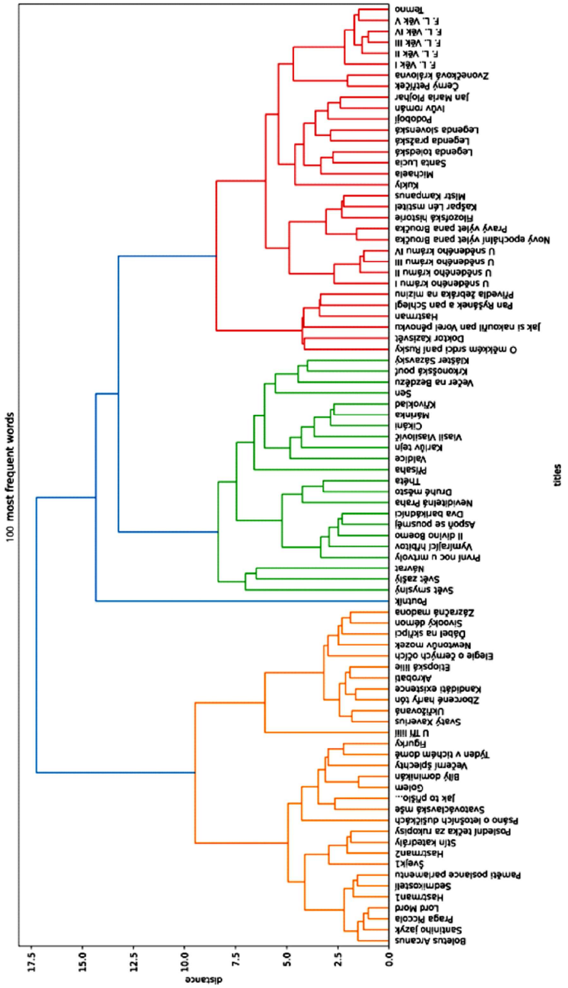


Fig. 7: Considering the size of some of Karel Hynek Mácha's prose, we calculated the relatedness between the texts with respect to the 100 most frequent words. From each text in the database, the 100 most frequent lemmas were selected and a common set of lemmas was created. For each lemma in this set, the relative frequency that the lemma has in each text was calculated and then a dendrogram was constructed. The distances between texts are expressed in the graph by the length of the y-axis.

Thus, as can be seen from this data arrangement, the potential analyses that such an essentially simple structure allows are many. For example, one can measure the percentage of word types, build one's own concordance or collocation searches, perform a number of statistical analyses of the text, e.g. measuring word richness, entropy, extensiveness, or concentration of texts (see Mistrík, 1968, pp. 40–52), detecting the so-called thematic concentration of texts (see Čech, 2016; Čech, Popescu, Altman, 2014, pp. 13–29), sentiment analysis[5] etc., or detecting the degree of similarity between texts using one of the stylometric methods (see Warmer-Colan, 2024). The following graph is an example of a so-called dendrogram, which shows the distances between titles in the whole existing database.

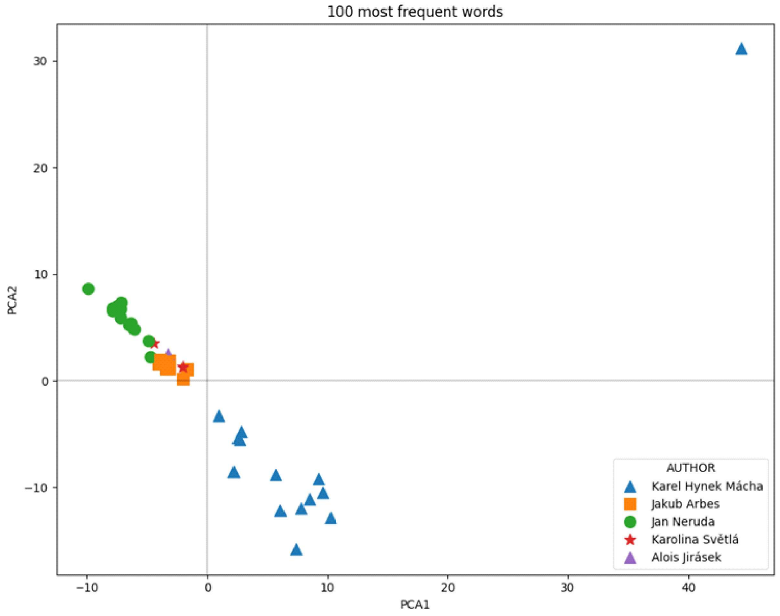


Fig 8: PCA analysis of a sub-corpus consisting of 40 selected texts (see Fig. 5).

[5] Sentiment analysis is a way of measuring the emotional load of texts, for example using special word lists or databases.

Similarly, the relatedness between texts can be modelled with the use of principal component analysis (PCA), which transforms multidimensional values into a two-dimensional representation that results in clusters of texts that are closest to each other. In the above graph, we can clearly observe the clusters of selected texts (see Fig. 8) distributed according to their affiliation to each author. The analysis was performed on the 100 most frequent words, but its criteria can be chosen according to different categories, e.g., word frequency in the different speech bands of the narratives, emotional load (so-called sentiment analysis), keywords, sentence lengths, types of n-grams, etc. As we can see in the graph, within the set of 40 texts that meet our criterion above, i.e. were first published in 1830–1880, the most distant clusters are prose works by Karel Hynek Mácha and Neruda's *Tales of the Lesser Town*. Especially in the case of Mácha, we can additionally observe their more pronounced internal diversification, which is due to the greater distances between the individual texts in Mácha's sub-corpus.

The criteria for working with such a database, which will of course be constantly updated and expanded, are similar to those for working with the *Gutenberg* library or the *DraCor* corpus; its user must have certain technical skills or experts who will be able to extract relevant information from such a dataset. For users who are used to standard ways of working with digital corpora, this project in particular also provides a web interface (https://korpusprozy.com) with a number of functionalities for corpus search. However, it is important to note that any web application with corpus tools is necessarily limited to certain tools. On the contrary, machine-readable and structured data allows for individual research and the development of specific tools for data mining. This can be observed in some foreign universities, which also engage in such practices directly in their teaching.[6] With the growing influence of digital humanities, there will be an increasing demand not only for special applications but also for specialized databases that allow researchers to conduct highly variable and independent research.

Translated from Czech by Josef Línek

References:

MISTRÍK, Josef. (1968). *Stylistics of the Slovak language*. Košice: Slovak Pedagogical Publishing House in Bratislava.

ČECH, Radek. (2016). *Thematic concentration of text in Czech*. Prague: Institute of Formal and Applied Linguistics.

ČECH, Radek; POPESCU, Ioan-Iovitz & ALTMAN, Gabriel. (2014). *Methods of quantitative analysis of (not only) poetic texts*. Olomouc: Palacky University in Olomouc.

DEFUS, A. (2024). What is stylometry? Available from: https://nauka.uj.edu.pl/aktualnosci/-/journal_content/56_INSTANCE_Sz8leL0jYQen/74541952/141176992.

WARMER-COLAN, A. (2024). Stylometry Methods and Practices. Available from: https://guides.temple.edu/stylometryfordh/home.

E-references

Project Gutenberg. Available from: https://archive.org/details/gutenberg.

DraCor. Available from: https://dracor.org.

Czech National Corpus. Prague: Institute of the Czech National Corpus FF UK. Available from: https://www.korpus.cz.

Corpus of Czech verse. Available from: https://versologie.cz/v2/web_content/corpus.php.

Literary Corpora. Available from: https://www.clarin.eu/resource-families/literary-corpora.

The Complete Corpus of Anglo-Saxon Poetry. Available from: https://sacred-texts.com/neu/ascp.

Korp – The Language Bank of Finland. Available from: https://korp.csc.fi/shibboleth-ds/index.html?https%3A%2F%2Fkorp.csc.fi%2Fkorp%2F%3Fsaved_params%3D1705922341337.

Distant Reading. Available from: https://distantreading.github.io/ELTeC.

Kramer. Available from: https://kramerius.nkp.cz/kramerius/Welcome.do;jsessionid=EDA92C4CCCFBB6ED5585E62C91C37BD3.

Literary Cartographic and Quantitative Models of Czech Novels from the 19th to 21st Century. Available from: https://korpusprozy.com.

[6] Cf. Statistical methods for studying literature using R at the University of Missouri-Kansas City (https://daedalus.umkc.edu/StatisticalMethods/index.html) or Mathew L. Jockers' Text Analysis with R for Students of Literature (2014).