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.

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:
```

---

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.

```
      (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/mode/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.
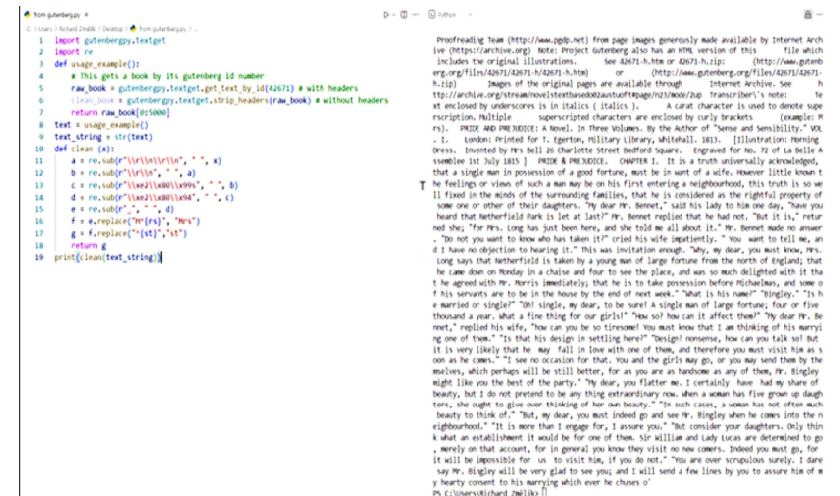

Fig. 2: Strings that are irrelevant to the text analysis are deleted from the text (see lines 11–17) in the left panel. The retrieval of a specific text is done via an ID that is identical to the eBook number in the Gutenberg database (cf. line 10 in Fig. 1 and line 5 here in the left panel)

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	Vlasil Vlasilovič	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 19st 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).
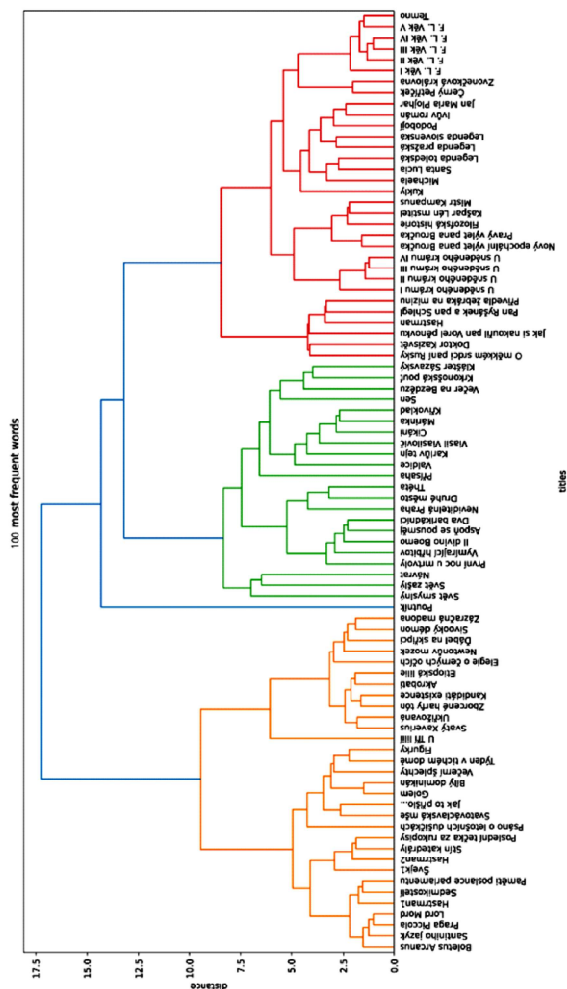
```
+----------------------+-------------------------------+----------+
| AUTHOR               | TITLE                         | 1. PUB   |
|----------------------+-------------------------------+----------|
| Karel Hynek Mácha    | Sen                           | 1832     |
| Karel Hynek Mácha    | 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]

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.

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.
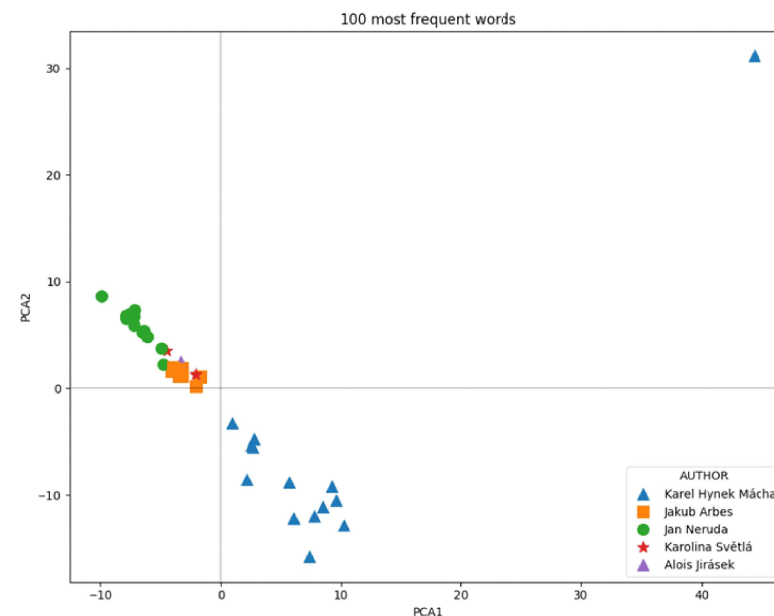

Fig 8: PCA analysis of a sub-corpus consisting of 40 selected texts (see Fig. 5).

[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).