Citation: Cinková, S., Hladká, B., Mírovský, J., & Archaimbault, S. (2024). Data storytelling around André Mazon's correspondence. *Colloquia Humanistica, 2024*(13), Article 3109. https://doi.org/10.11649/ch.3109

Colloquia Humanistica 13 (2024)
Digital Studies of Culture in Central Europe
Article No. 3109
https://doi.org/10.11649/ch.3109

Silvie Cinková

Charles University, Prague
https://orcid.org/0000-0003-4526-3915
cinkova@ufal.mff.cuni.cz

Barbora Hladká

Charles University, Prague
https://orcid.org/0000-0003-4950-4587
hladka@ufal.mff.cuni.cz

Jiří Mírovský

Charles University, Prague
https://orcid.org/0000-0003-2741-1347
mirovsky@ufal.mff.cuni.cz

Sylvie Archaimbault

Sorbonne Université, Paris
https://orcid.org/0000-0002-7869-5098
sylvie.archaimbault@sorbonne-universite.fr

Data Storytelling Around André Mazon's Correspondence

Abstract

We present a data story, the central concept of our international one-semester data analytics course for students of social sciences and humanities. Namely, we demonstrate the four stages of the data lifecycle – gathering, analyzing, annotating, licensing & sharing – using the multilingual

Publisher: Institute of Slavic Studies, Polish Academy of Sciences
Editor-in-chief: Jolanta Sujecka
Conception and academic editing: Robert Kulmiński and Ondřej Daniel

correspondence collection of French Slavist André Mazon and tools for data visualization (Tableau Public), text transcription (Transkribus, Pero), building and searching text corpora (TEITOK, Corpus Query Language) and natural language processing (UDPipe).

Keywords: teaching, data story, correspondence archive, data analysis, natural language processing.

Introduction

Czech universities have been strongly involved in both national and international research projects in the field (including large research infrastructures such as MetaNet, CLARIN, and DARIAH), but have taken very long to explicitly adapt and institutionalize the digital humanities in teaching. It began with short-term events such as *Humanities Going Digital* at Palacký University, which was set up in collaboration with the University of Pécs and Universidade Nova de Lisboa as a COVID-enforced substitution for Erasmus (2020), as well as a shorter training school (Autumn School of Digital Humanities, 2022).

Charles University, the largest in the country, organized several major events, such as a natural language processing (NLP) tutorial in the digital humanities within the Visegrad Fund-financed project *Training Digital Scholars: Knowledge Exchange Between V4 and Austria*, Series (2018–2020),[1] and the Text Encoding Initiative Public Day dedicated to methodologies in the digital humanities (supported by Marie Curie Actions in 2017),[2] and, most recently, a data and annotation summer school within the EC-funded *Horizon 2020 Computational Literary Studies Infrastructure* project (2022),[3] to name just a few.

According to a survey in the 2018/2019 academic year, Czech universities were teaching virtually no courses explicitly related to the digital humanities (Cinková et al., 2022), although many historical, philological, linguistic and archive/librarianship curricula across the country taught elementary skills in programming, statistics, digital editions, natural language processing, and markup languages. Since then, the digital humanities have gained

[1] https://ufal.mff.cuni.cz/training-digital-scholars-knowledge-exchange-between-v4-and-austria

[2] https://www.ff.cuni.cz/2017/01/methodologies-digital-humanities-text-encoding-initiative-public-day-2

[3] https://campus.dariah.eu/resource/events/cls-infra-training-school-on-data-and-annotation

much more ground. For instance, Palacký University in Olomouc now runs a dedicated BA curriculum called Linguistics and Digital Humanities (since 2021, as a result of a bilateral project with the University of Wrocław).[4] At Charles University, the digital humanities are not yet taught within a dedicated curriculum. However, its Faculty of Arts established a virtual Department of Digital Humanities in 2022, most likely with the intent to make it responsible for such a program. So far, 12 courses at the Faculty of Arts contain the keyword "digital humanities" in their title or annotation, related to geospatial and geostatistical analysis, data science for linguists and historians, and multimedia. Masaryk University in Brno, the country's second-largest university, offers 13 such courses, mostly associated with library and information science, archive and history studies, and philology.

A recent survey – @CLS INFRA D4.1 Skills Gap Analysis – has revealed the current distribution of self-attested practical text-processing skills as well as the corresponding education supplies among digital humanities researchers of all academic stages (van Rossum & Šeļa, 2022). The most desired skills were research design, corpus building, classification, text modeling, feature engineering and data visualization, whereas the related skills of annotation and corpus analysis scored surprisingly low. This implies that students fail to identify them as prerequisites for the skills they want to acquire.

The conclusion is that students would better learn text processing methods when embedded in realistic workflows rather than individually. We therefore present Data Stories, the central concept of our international one-semester data analytics course for students of social sciences and humanities.[5] Our course presents four stages in the usual data lifecycle – gathering, analyzing, annotating, licensing & sharing – in realistic scenarios with guided hands-on tasks. The most realistic scenario is the multilingual correspondence collection of French Slavist André Mazon (1881–1967) and its metadata in a spreadsheet. Among the scanned letters, students are instructed to select items with the highest processing priority, visualize diverse characteristics in interactive dashboards (Tableau Public), and argue for their decisions. Then they transcribe the original scans with Transkribus/Pero. They load the texts into a corpus manager (TEITOK), formulate a research question, operationalize it in lexico-morpho-syntactic terms and extract information

[4] https://studium.upol.cz/Catalog/StudyPrograms?type=Bachelor#year=2023&globalId=45406&maior=7165

[5] https://ufal.mff.cuni.cz/courses/npfl134

with the Corpus Query Language and PML-TQ, using the linguistic markup they could add by running UDPipe. In addition, they can manually link entities to WikiData and VIAAF. While trying out and using the tools, students learn to formulate a question, present their methodological decisions, and report and discuss the results.

The paper is organized into four sections corresponding to the main stages of the usual data lifecycle, illustrated with a case study on the Mazon collection.

Data Gathering

The Mazon collection has been maintained by the Sorbonne University within the NUMERISLAV project as one of 25 archival collections related to French Slavic studies in the 20th century. It mostly comprises manuscripts and correspondence from the scholars' personal estates, donated by their descendants.

Many of the French-based Slavists were immigrants from Slavic countries suffering from the political turbulence occurring throughout the century, including some of André Mazon's correspondents. Many had been public persons in their home countries' politics or diplomacy, or at least part of those social circles. The NUMERISLAV project has thus become an important interdisciplinary research center not only for Slavic philology but also history and political science.

The Mazon collection contains about 3 thousand scanned letters, virtually all addressed to André Mazon. In terms of machine readability, the letters are images not yet recognized as texts. The images typically render one page or one double page of handwritten text in a separate jpg file. Each letter is stored in an individual folder. The folder name serves as the unique ID of each letter.

The catalog of 2,112 letters from the Mazon collection, the so-called metadata file, is a csv table, where each letter is represented by one row. It records the letter's ID, date and location of writing, language, author and recipient, as well as the document type (e.g. typed letter, written postcard). The NUMERISLAV archivists also linked some authors to VIAF.[6] To this information, we added the GPS coordinates.

[6] https://viaf.org

Data Analysis

Data analysis is a process of deeper understanding of a task by taking a statistical view of the data. It consists of (1) data inspection, which includes checking attributes and their values (e.g. the attribute Language having 14 different values in the Mazon catalog), (2) exploratory analysis to summarize the characteristics of the data (e.g. the distribution of languages in the catalog), and (3) statistical testing to make inferences from the data.

For the Mazon collection, it would appear most straightforward to explore such a large text collection with text mining techniques to extract keywords, topics, named entities, etc. Nevertheless, this is not an option for the Mazon collection, since most of the letters are not machine-readable texts but images that first have to be transcribed with a handwritten text recognition (HTR) tool based on machine learning techniques. This is going to take a long time if the tool is not trained on similar data (see Section 3). The students are thus assigned their first task: to explore the metadata file and argue which letters to prioritize, choosing one of two goals, i.e. to train a machine learning model for the HTR tool, or conduct research on a very prominent topic. Possible arguments for either: there are many letters by one given author in one given language (i.e. just one HTR model to train), and/or the author kept corresponding with Mazon for their lifetime, the author is very famous or interesting, or the letters from a location in a given period can reflect some interesting historical moments (e.g. the onset of the Russian Revolution or a world war). Obviously, such data analysis tasks do not have one correct solution; their true purpose is to teach students foundational data literacy, i.e. simple descriptive statistics, filtering and joining of tabular data, and visualization/reporting. The tool of choice is Tableau Public.[7] It displays visualizations on the public web. Its conceptual building blocks are worksheets, dashboards, and data stories. One worksheet corresponds to one visualization based on one imported spreadsheet or a joined database of spreadsheets. Tableau creates its own image of the imported data, i.e. it never modifies the original. Worksheets are combined into dashboards. Dashboards can contain other content too, such as media or text. A presentation of different dashboards is called a data story. It creates a navigation panel with comments to each dashboard.

[7] https://public.tableau.com, a free module on the Tableau business intelligence platform, accessed 2023–01–06

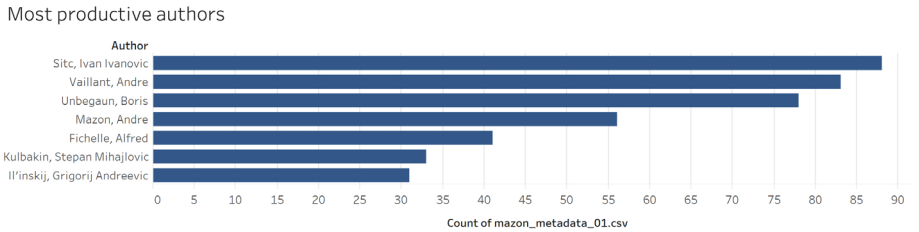

Figure 1. Bar plot of letters by the authors who wrote at least 30 letters.*

Tableau has an impressive offer of plots and diverse tabular elements. Each of these elements can be enriched with an interactive filter for the user. The user can then zoom in on data points of their individual interest. Fig. 1 shows a bar plot for letters by individual authors, where an interactive slider extracts only authors who are represented by at least 30 letters in the collection.

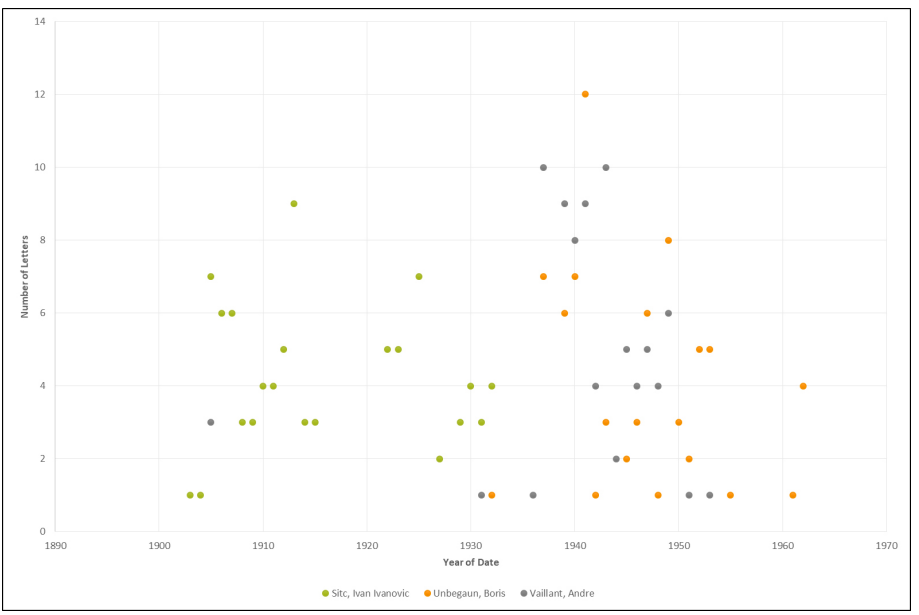

Figure 2. Dot plot of letters by the three most productive authors.

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

The plots operate with axes, shapes, size, labels and colors, as well as an interactive tooltip (pop-up text on data points). Fig. 2 displays the contributions of the three most productive authors on a timeline: Ivan Ivanovič Šitc (1874–1942), André Vaillant (1890–1977) and Boris Unbegaun (1898–1973). By 1915, Šitc had written Mazon 49 letters since 1905, and he had 33 letters to write him to total 88 letters, with his last letter arriving in 1932. In the collection, Šitc clearly represents a person who had good contact with the young Mazon. Vaillant's correspondence with Mazon lasted 48 years, starting as early as 1905 and ending in 1953, with 83 letters. The collection reflects only sporadic contacts in the first two decades, suddenly exploding as if under the threat of war, similarly to Unbegaun who wrote Mazon 78 letters in total.

Another plot (Fig. 3) shows the spatial distribution of the correspondence between 1903 and 1967.

These are examples of worksheets the students learn to create in order to argue for interesting authors or spatial or temporal slices of data to prioritize on the basis of the metadata. At the end of our course, the students presented their own research to the class, in the form of a Tableau dashboard or a data story. We encouraged them to self-study the Mazon collection, focusing

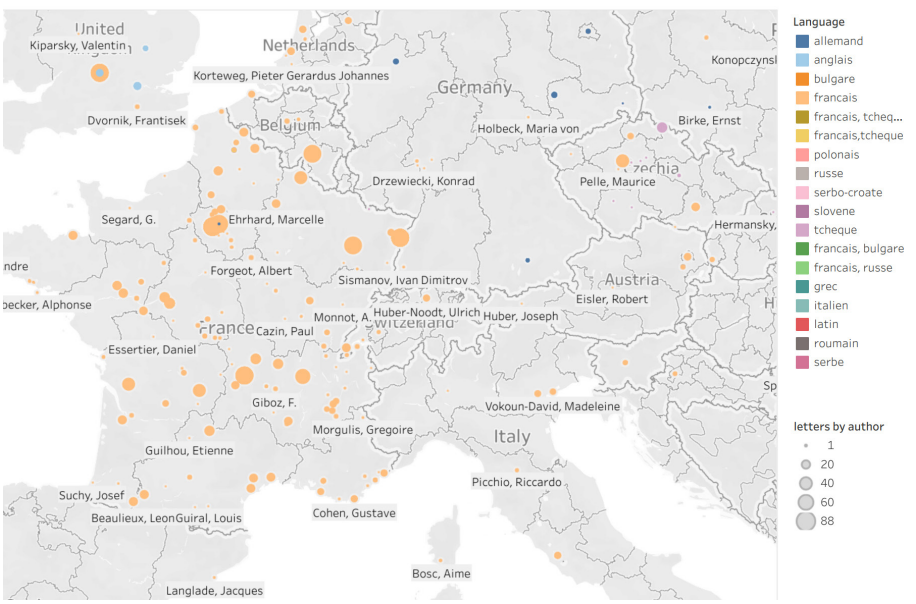

Figure 3. Spatial distribution of the correspondence between 1903 and 1967.

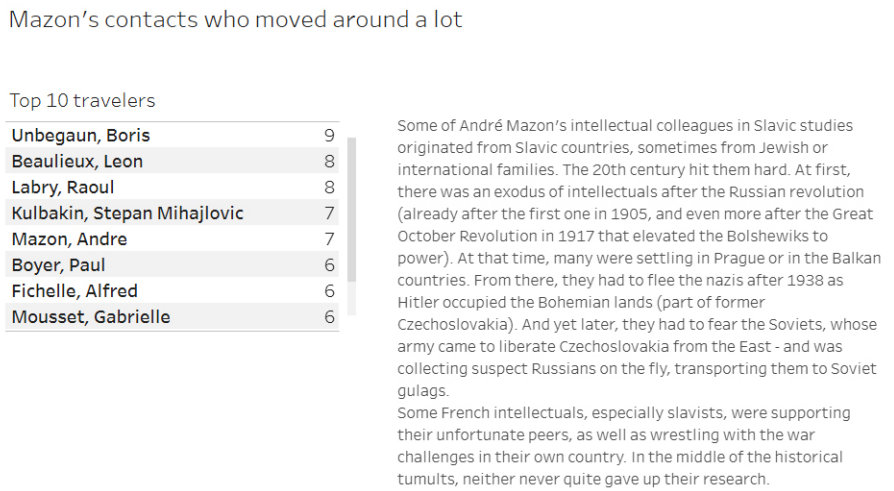

Figure 4. Mazon's contacts who moved around a lot.

on the "research matter" itself on the one hand and the technical details of the resulting Tableau objects on the other. The research questions were typically presented like this:

"I wonder which pen friends of Mazon traveled a lot. I operationalize it as the number of different places of writing, with all the limitations involved. Then I will choose one author and describe his life itinerary, as reflected by the letters preserved in the Mazon collection. To begin with, I make a pivot table with the authors and places in columns, and let Tableau calculate the unique counts. Then I employ a filter to extract the top 10 authors in terms of unique places. I also display the places from which these people wrote their letters. The biggest traveler is Boris Unbegaun, with nine places (see Fig. 4), and I would like to know how he moved around. This is what I have done: a map with the places where his letters were written, connected with paths, and different years in different colors. The letters are arranged chronologically. When a user hovers their mouse over a point, they see the name of the place, the year, and again the chronological rank index of the given letter. They can zoom in and out and pan the map, so that, for example, they can follow the individual letters written in one place. Technically, this plot is called a *spider map* or a *path map* and is typically used for airline routes, wind directions, and shared bike routes. To do this,

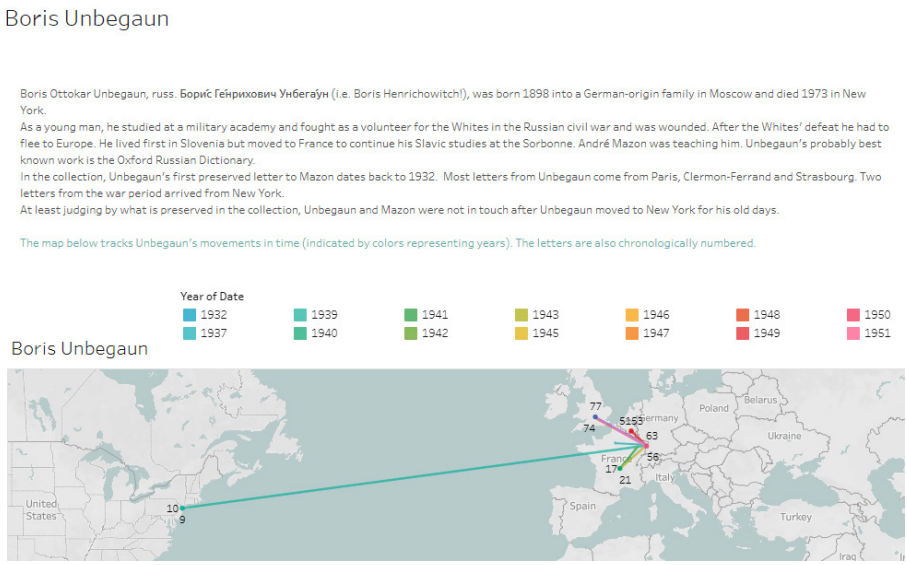

Figure 5. Data story of Boris Unbegaun's travels.

I needed the following information for each letter: author, place, its GPS longitude and latitude coordinates, and I also needed to generate a column with the chronological rank for each place grouped by author. Without it, the points were connected somewhat randomly, and I could not make Tableau understand that they should connect directly by dates.

I searched the Tableau tutorials and found the missing information that I needed a separate column with explicit ranks. This seemed cumbersome in Tableau, so I did this quickly in Google spreadsheets and uploaded this file back to Tableau. First, I created a map with longitude as "columns" and latitude as "rows". Then I filtered Author to extract just Unbegaun. I mapped YEAR(Date) as a Dimension on color and the rank on Path. Path appeared as an option when I selected "Line" in the "Marks" menu. I also mapped the rank on "Text", so I could see the ranks as labels for control. Eventually, I gathered my notes from the web and added them to the worksheet to create a dashboard, see Fig. 5. Together with the dashboard in Fig. 4, this map could form a nice data story."

Data Annotation

So far, we have only worked with meta information about the Mazon collection. The next part (of the course) focuses on annotating and using the content of the letters. Data annotation is the process of labeling data to make it easier for computers to understand and interpret. Given the nature of the Mazon correspondence collection (scanned letters), the first step before any annotation could start must be the transcription of the handwritten documents to machine-readable text files. For this task, two publicly available web tools are used within the course: PERO[8] for Czech documents and Transkribus[9] for all other languages. Out of all the scanned documents in the Mazon collection, a few hundred in various languages have been selected and compiled into several dozen bundles of about three to six pages, depending on the extent and readability of the content. These bundles have – depending on the language – been uploaded into the appropriate transcription system. The students can choose their preferred language and then they are assigned a single bundle to process as homework.

Two parallel classes – one for PERO, one for Transkribus – are dedicated to introducing the students to the process of using transcription tools. While the homework is designed to start with documents already uploaded in the tool, in the introductory classes the process is presented in its entirety, i.e. from uploading the scanned documents to the system, setting up the recognition process, marking the layout, through automatically processing the documents and manually correcting the results, to exporting the transcribed data to local text and XML files. While the text format suffices as a representation of the content of the documents, the XML format preserves the interconnection of the text with the original scanned documents.

Once the documents are transcribed and available in text and XML formats, they need to be further processed to allow all kinds of linguistic analyses. From various existing natural language processing (NLP) frameworks, we have decided on Universal Dependencies (UD; de Marneffe et al., 2021), which is "a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages" (*Universal Dependencies*, n.d.). In a search for a suitable tool, we also sought to satisfy two constraints: (i) to preserve

[8] https://pero-ocr.fit.vutbr.cz
[9] https://readcoop.eu/transkribus

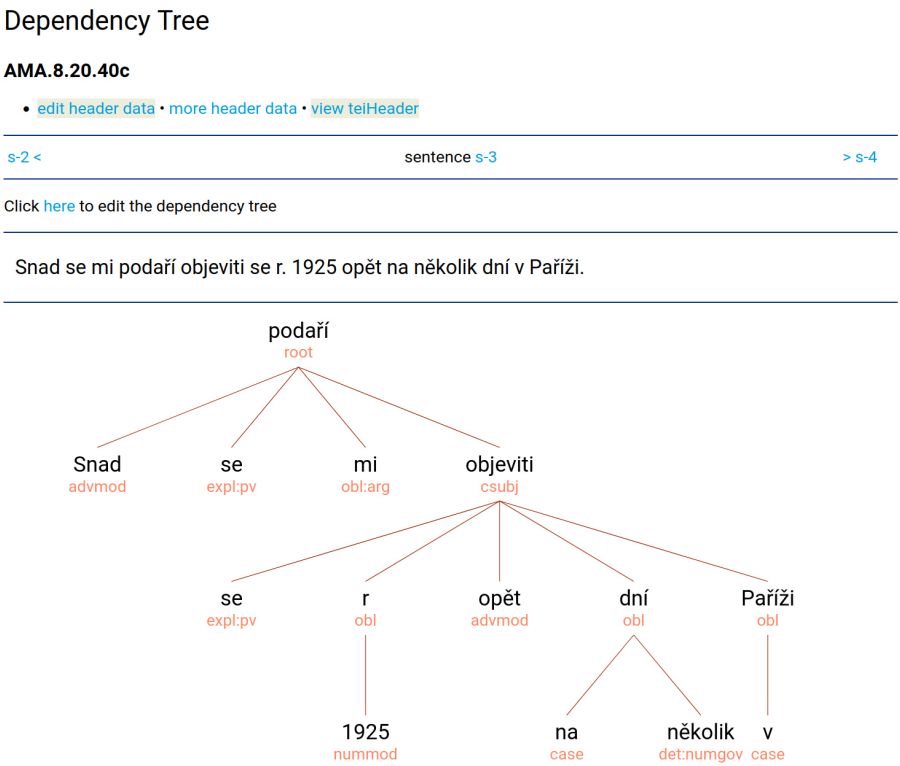

Figure 6. TEITOK interface: a sentence and its syntactic dependency tree.

the interconnection of the texts with the scanned documents, and (ii) to avoid the need to use any programming in the course. All the requirements are met by TEITOK (Janssen, 2016), "a web-based system for viewing, creating, and editing corpora with both rich textual mark-up and linguistic annotation" (*TEITOK Corpora at LINDAT*, n.d.).

Two classes of the course are then dedicated to introducing Universal Dependencies and TEITOK. The students are taught to upload the scanned documents along with their transcriptions into the system, organize them into a corpus, run an automatic analysis in the UD framework, inspect the results and manually fix errors. See Fig. 6 as an example of a Czech sentence displayed in TEITOK along with its automatically parsed dependency tree. Extensive time is devoted to teaching how to search the data using two search engines inbuilt in TEITOK: CQL for a search

in unstructured data, and PML Query Language (PML-TQ; Pajas & Štěpánek, 2009) for a complex search in syntactically parsed data. This constitutes another piece of homework for the students, as they are asked to come up with their own research question on the (part of) Mazon data in TEITOK, design and perform a required search on the data, and prepare a presentation of the results.

Data Licensing & Sharing

When authors make their data available to others, they should observe two rules: (1) always make the data available under a license, so that it is clear to any person wishing to access and use them who owns the data, and on what terms they may be used; (2) make the data available under the most open license possible, which allows the widest possible scope for re-use and redistribution. Data should be made available under an open license, unless there is a good reason to license them on a more restrictive basis, e.g. to prohibit commercial re-use of data in which a commercial partner has an interest.

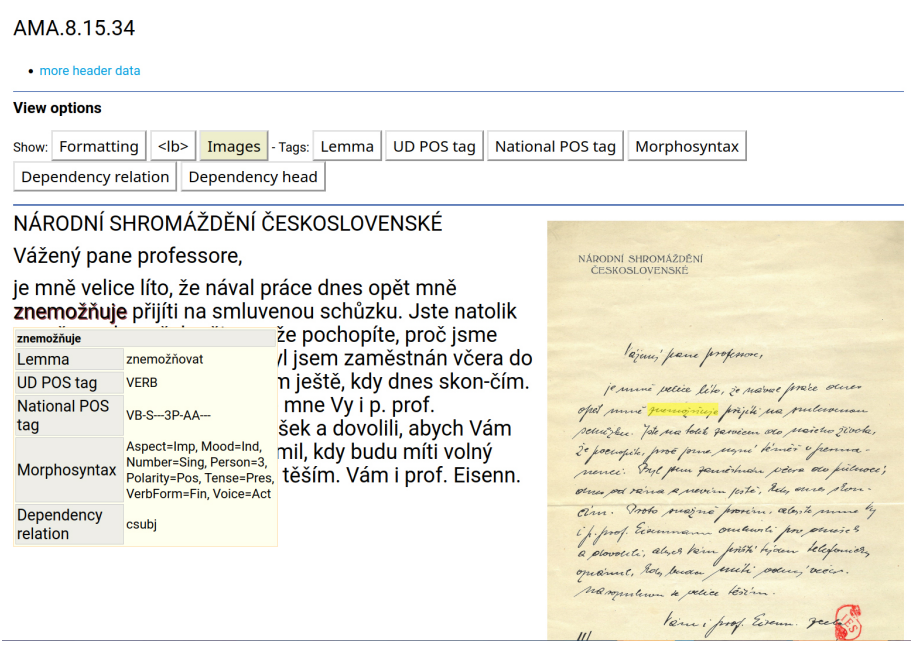

Figure 7. TEITOK interface: a text next to its original scan with a highlighted word.

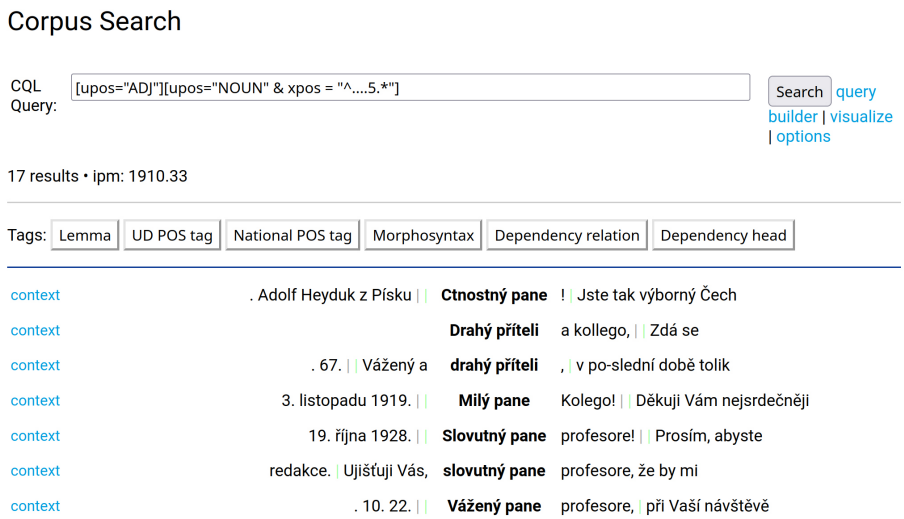

Figure 8. CQL search in TEITOK.

The complete André Mazon correspondence has not been published yet. Therefore we make a sample of 29 letters that we have at our disposal for teaching purposes accessible in TEITOK. The Mazon collection in TEITOK serves not only as a demonstration and teaching environment for the students of the course but is also publicly available to researchers worldwide. It is hosted at the LINDAT/CLARIAH-CZ repository[10] and accessible via Authentication and Authorization Infrastructure (AAI).[11] The TEITOK interface offers four ways of exploring the data:

1) Browse and view individual documents based on their tags: language (Czech, English, French, German, Russian) and document type (handwritten letter, handwritten postcard, typed letter). A document can be displayed next to its original scan; hovering above a word opens a pop-up window with Universal Dependencies-related information (form, lemma, tag, dependency relation etc.) and highlights a corresponding area in the original scan. In Fig. 7, a Czech-written letter is displayed as a text next to its scan, with detailed information about the highlighted word "znemožňuje" ("prevents from").

[10] https://lindat.cz/services/teitok-live/mazon/index.php
[11] https://elixir-europe.org/platforms/compute/aai

Figure 9. Facsimile search in TEITOK.

2) The collection can be searched using the aforementioned CQL, allowing complex search patterns on the properties of individual words or their sequence to be set; for example, we can search for adjectives immediately followed by a noun in vocative, thus looking for typical letter openings. Pieces of documents matching the query are then displayed in the KWIC (keyword in context) format. See Fig. 8 for an example query and a part of the result with highlighted matches: "Ctnostný pane" ("Virtuous sir"), "Drahý příteli" ("Dear friend"), "Milý pane" (lit. "Nice sir"), "Slovutný pane" ("Famous sir") and "Vážený pane" ("Esteemed sir").
3) CQL can also be used in the Facsimile search, which displays the matching words directly next to the corresponding segments in the original scans; see Fig. 9 for the result of a search for the word "příteli" ("friend" e.g. in vocative).
4) For more complex queries, TEITOK offers the PML-TQ (Prague Markup Language – Query Language), which allows search patterns to be set not only on individual words and their sequence but also on the dependency structure of the sentence. The results of a PML-TQ query can be viewed individually or processed further with a powerful system of output filters, which allow for complex summarization and tabular representation of the results. It is possible to obtain, for example, the distribution of groups of dependency functions of a given verb complementation.

Reception and Feedback by the Students

The course was designed within the 4EU+ program, the data-providing partner being the Sorbonne. Hence we expected most students to be Sorbonne's Slavistics students. Eventually, administrative hurdles at the Sorbonne hampered its own students' participation in the course. Most of our students originated from other 4EU+ partner universities, from Milan in particular. Their educational profile was closer to social sciences than Slavic studies. Hence they were clearly focusing on the methods (data wrangling, plotting, machine learning) rather than the field of knowledge itself. Most of them had no programming skills and embraced an introductory course focused on use cases and interactive tools. At the same time, their practical experience with off-the-shelf tools made them aware of these tools' limitations, and they repeatedly expressed their motivation to learn a programming language.

Conclusion

We have presented a data story on a correspondence archive as a case of research-based learning (Ifenthaler & Gosper, 2014) in the digital humanities, providing a safe environment for confronting students with the uncertainty of research (What are the best prioritization criteria for processing certain documents over others? How well do NLP tools actually perform on a given domain?) and the inevitability of ad hoc design decisions and their implications (manual revisions of automatic Handwritten Text Recognition, corpus queries), as well as the limitations of interactive tools (e.g. reproducibility issues). The students learned how to navigate tabular metadata, how to extract content from unstructured textual data, and how to properly share their data in a community repository to make their research reproducible.

References

Cinková, S., Škvrňák, J., & Škvrňák, M. (2022). Výuka digitálních humanitních věd na českých veřejných vysokých školách podle latentní sémantické analýzy. In R. Hladík (Ed.), *Digitální obrat v českých humanitních a sociálních vědách* (pp. 367–409). Karolinum.

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

Ifenthaler, D., & Gosper, M. (2014). Research-based learning: Connecting research and instruction. In M. Gosper & D. Ifenthaler (Eds.), *Curriculum models for the 21st century: Using learning technologies in higher education* (pp. 73–89). Springer. https://doi.org/10.1007/978-1-4614-7366-4_5

Janssen, M. (2016). TEITOK: Text-Faithful Annotated Corpora. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 4037–4043). European Language Resources Association (ELRA).

Pajas, P., & Štěpánek, J. (2009). System for querying syntactically annotated corpora. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations* (pp. 33–36). Association for Computational Linguistics. https://doi.org/10.3115/1667872.1667881

TEITOK Corpora at LINDAT. (n.d.). https://lindat.cz/services/teitok/index.php

Universal Dependencies. (n.d.). https://universaldependencies.org

van Rossum, L. M., & Šeļa, A. (2022). *CLS INFRA D4.1 Skills gap analysis*. Zenodo. https://doi.org/10.5281/zenodo.6401858

Příběh dat korespondence André Mazona

Představujeme datový příběh, ústřední koncept našeho mezinárodního jednosemestrálního kurzu datové analýzy pro studenty společenských a humanitních věd. Konkrétně přibližujeme čtyři fáze životního cyklu dat – shromažďování, analýzu, anotaci, licencování a sdílení – na vícejazyčné sbírce korespondence francouzského slavisty André Mazona a nástrojích pro vizualizaci dat (Tableau Public), přepis textu (Transkribus, Pero), vytváření a prohledávání textových korpusů (TEITOK, Corpus Query Language) a zpracování přirozeného jazyka (UDPipe).

Klíčová slova: výuka, datový příběh, archiv korespondence, analýza dat, zpracování přirozeného jazyka.

Historia danych w korespondencji André Mazona

W artykule prezentujemy historię danych, główną ideę naszego międzynarodowego, semestralnego kursu analizy danych dla studentów nauk społecznych i humanistycznych. Posługując się wielojęzycznym zbiorem korespondencji francuskiego slawisty André Mazona, przybliżamy w szczególności cztery fazy cyklu życia danych – gromadzenie, analizę, anotację, licencjonowanie i udostępnianie. Wykorzystujemy przy tym narzędzia do wizualizacji danych (Tableau Public), transkrypcji tekstu (Transkribus, Pero), tworzenia i przeszukiwania korpusów tekstowych (TEITOK, Corpus Query Language) oraz przetwarzania języka naturalnego (UDPipe).

Słowa kluczowe: nauczanie, historia danych, archiwum korespondencji, analiza danych, przetwarzanie języka naturalnego.

Przekład z języka czeskiego
Robert Kulmiński

Note

Silvie Cinková, Charles University, Prague, Czechia.
cinkova@ufal.mff.cuni.cz
https://orcid.org/0000-0003-4526-3915

Barbora Hladká, Charles University, Prague, Czechia.
hladka@ufal.mff.cuni.cz
https://orcid.org/0000-0003-4950-4587

Jiří Mírovský, Charles University, Prague, Czechia.
mirovsky@ufal.mff.cuni.cz
https://orcid.org/0000-0003-2741-1347

Sylvie Archaimbault, Sorbonne University, Paris, France.
sylvie.archaimbault@sorbonne-universite.fr
https://orcid.org/0000-0002-7869-5098

This article was created with equal participation of all authors in conceptual work, excerption of material, and analysis development.
Preparation of the paper was funded by institutional sources of the respective authors and by the LINDAT/CLARIAH-CZ project of the Ministry of Education of the Czech Republic (project LM2023062).
No competing interests have been declared.

Publication History

Received: 2023-09-29, Accepted: 2024-05-21, Published: 2024-12-23