Multilingual Analysis and Visualization of Bibliographic Metadata and Texts With the AVOBMAT Research Tool

ABSTRACT
The objective of this paper is to introduce the workflow of the AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) multilingual research tool, which enables researchers to critically analyse bibliographic data and texts at scale with the help of data-driven methods supported by Natural Language Processing (NLP) techniques. This exploratory tool offers a range of dynamic text and data mining tasks and provides interactive parameter tuning and control from the pre-processing to the analytical stages. It can pre-process, analyse and (semantically) enrich a vast number of texts and metadata in several languages due to its scalable infrastructure. The implemented analytical and visualization tools provide close and distant reading of texts and bibliographic data. It combines bibliographic data and NLP research methods in one integrated, interactive, user-friendly web application, allowing users to ask complex research questions.

CORRESPONDING AUTHOR:
Róbert Péter
Institute of English and American Studies, University of Szeged, Szeged, Hungary
robert.peter@ieas-szeged.hu

KEYWORDS:
data analysis; bibliographic data; content analysis; linguistic analysis; modelling; databases; multilingualism

TO CITE THIS ARTICLE:
THE PAPER

The aim of this paper is to introduce the workflow of the AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) multilingual research tool. (Péter et al., 2020; Péter et al., 2022).

1. UPLOADING THE CORPUS

Users can upload metadata and texts in several formats: Zotero collections in CSV and RDF formats and EPrints (library) repositories as XML files (metadata or metadata with links to the full texts). AVOBMAT can also import full texts, for example, by uploading a zip file of documents along with a CSV of the metadata. Documents from external databases can be imported by providing URLs to the full texts in the CSV. It can process texts in several formats since the Apache Tika library converts them to plain text.

2. CLEANING THE CORPUS

AVOBMAT provides several options for cleaning the text corpus. For example, users can

• remove non-alphabetical tokens (e.g. of OCR-ed texts);
• upload a list of words and replace words (e.g. synonyms) and characters;
• make use of regular expressions.

A context filter is implemented to keep the context of a keyword or keywords and remove all other parts of the document.

3. CONFIGURING THE PARAMETERS

Users can create different configurations for each analysis where the outcome depends on the language of the texts. There are two ways to assign a language to a document: researchers can manually select a language for the full dataset (52 languages) or choose the automatic language detection option. As for the latter, the system will choose a language independently for each document. Based on the language, it offers stopword and punctuation, filtering drawing on the spaCy library, and lemmatization (SpaCy Models and Languages). Extra stopword and punctuation lists can also be added. SpaCy language models are used for lemmatization, with LemmaGen models being used for languages not supported by spaCy (Juršic et al., 2010).

The following pre-processing options are implemented:

• choose spaCy language model (small, large or transformer);
• make text lowercase;
• remove numbers;
• set minimal character length.

The metadata enrichment includes the identification of the gender of the authors (male, female, unknown gender or without author) and automatic language detection. Users can also upload a list of male and female first names, supplementing and replacing the ones found in the dictionaries of the programme.

As for topic modelling, the user also has the option to separate the documents into sections of equal size. Users can specify the so-called window length for certain lexical diversity analyses (MSTTR, MATTR).

4. VALIDATING THE SETTINGS

AVOBMAT cleans and pre-processes a small sample of the uploaded database where the user can check if the set parameters are appropriate. The settings can be saved in a template if the configuration is acceptable. If the parameters need to be fine-tuned, the user can start the cleaning and configuration process again. AVOBMAT identifies missing values and gaps in the metadata.
5. FILTERING THE CORPUS

The user can search and filter the metadata and texts in faceted, advanced and command-line modes and perform all the subsequent analyses on the filtered dataset (Figure 1). The NLP analyses of the documents semantically enrich the metadata. For example, the recognized named entities such as person appear in all types of searches and the user can search for (disambiguated) named entities in 16 languages. The tool supports fuzzy and proximity searches.

6. INTERACTIVE METADATA ANALYSIS

Having filtered the uploaded databases and selected the metadata field(s) to be explored (Figure 2), the user can, among other actions,

- analyse and visualize the bibliographic data chronologically in line and area charts in normalized and aggregated formats (Figure 4);
- create an interactive network analysis of the metadata fields (Figure 3);
- make pie, horizontal and vertical bar charts.

![Figure 1 AVOBMAT graphical interface.](image)

![Figure 2 Interactive metadata visualization setting.](image)
7. INTERACTIVE CONTENT ANALYSIS

The following options are available for interactive content analysis.

7.1. N-GRAM VIEWER

This diachronic analysis of texts shows the yearly count of the specified n-grams generated at the pre-processing stage in aggregated and normalized views (Figure 5).

7.2. FREQUENCY ANALYSIS

Frequency analyses and word clouds can be efficient tools to highlight the prominent terms in a corpus. The significant text analytical tool shows what differentiates a subset of the documents from others using four different metrics (e.g. Chi square) (Manning et al., 2009; Rudi and Vitányi, 2007; see Significant text aggregation). The TagSphere analysis enables users to investigate the context of a word by creating tag clouds showing the co-occurring words of a specified search term within a specified word distance (Figures 6 and 7) (Jänicke and Scheuermann, 2017). Words can be interactively removed from the clouds. Bar chart versions of the analyses present the applied scores and frequencies.
Figure 5 N-gram viewer. Distribution of “katolikus egyház” [Catholic church], “református egyház” [Reformed church], “evangélikus egyház” [Lutheran church] in the Délmagyarország daily newspaper, 1911–2009.

Figure 6 TagSphere analysis. Dan Brown’s novels. Keyword: god, word distance: 4 (shown in different colours), minimum word frequency: 7, lemmatized texts, stopwords removed.

Figure 7 The same TagSphere analysis as in Figure 6. Bar chart view with statistical data.
7.3. LEXICAL DIVERSITY

AVOBMAT calculates the lexical diversity of texts according to eight different metrics: Type-token ratio (TTR), Guiraud, Herdan, Mass TTR, Mean segmental TTR, Moving average TTR, Measure of Textual Lexical Diversity and Hypergeometric distribution Diversity (Figure 8) (Covington and McFall, 2010; McCarthy and Jarvis, 2010; Torruella and Capsada, 2013).

7.4. KEYWORD-IN-CONTEXT

The keyword-in-context function supports the close reading of texts (Figure 9).

7.5. TOPIC MODELLING

The Latent Dirichlet Allocation function calculates and graphically represents topic models (Blei et al., 2003). It shows the most relevant words and most relevant documents related to each topic, visualizes the distribution of these topics chronologically, highlights the correlation of different topics and exports the results in various formats (Figures 10 and 11). It has the following parameters: the minimum number of occurrences of words, the number of topics and iterations, per-document topic distribution (alpha), and per-topic word distribution (beta) parameters. Users can interactively remove stopwords.
7.6. PART-OF-SPEECH TAGGING

AVOBMAT identifies the part-of-speech tags currently in 16 languages by using the spaCy language models. It produces different interactive visualizations and statistical tables of the results (Figures 12 and 13).
7.7. NAMED ENTITY RECOGNITION, DISAMBIGUATION AND LINKING

It identifies named entities such as persons and places currently in 16 languages. The number and type of named entities differ by language, as seen in Table 1. AVOBMAT creates different statistical tables and visualization of these entities. The latter are displayed in full-text view. As for the English language, it disambiguates the entities and links them to Wikidata, VIAF and ISNI (Figure 14).

<table>
<thead>
<tr>
<th>Language model</th>
<th>Person</th>
<th>Organization</th>
<th>Location</th>
<th>Miscellaneous</th>
<th>Language</th>
<th>Work of art</th>
<th>Geopolitical entity</th>
<th>National or religious group</th>
<th>Date</th>
<th>Ordinal number</th>
<th>Product</th>
<th>Money</th>
<th>Infrastructure</th>
<th>Cardinal number</th>
<th>Event</th>
<th>Law</th>
<th>Percent</th>
<th>Period</th>
<th>Movement</th>
<th>Phone</th>
<th>Pet name</th>
<th>Title</th>
<th>Title affix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Danish</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Dutch</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>English</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>French</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>German</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Greek</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Italian</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Japanese</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Lithuanian</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Hungarian</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Norwegian Bokmål</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Polish</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Romanian</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Multi-language</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
8. EXPORT RESULTS, CONFIGURATIONS AND PUBLICIZE DATABASES

The reproducibility and transparency of the experiments and results using the tool are enhanced by the ability to import and export the parameter settings in JSON format. The users can create templates for the pre-processing and analytical functions on the graphical interface. The tabular statistical data and visualizations of the performed analyses can be saved in PNG and different CSV formats, including a document-topic graph file for Gephi in case of topic modelling. The latter enables researchers to use the generated data in other software. Users can share and make their databases public.

FUNDING INFORMATION

The creation of the AVOBMAT software was partially funded by the EFOP-3.6.1-16-2016-00008, EFOP-3.6.3-VEKOP-16-2017-0002, 2019-1.2.1-EGYETEMİ-ÖKO-2019-00018 and the Humanities and Social Sciences Cluster of the Centre of Excellence for Interdisciplinary Research, Development and Innovation of the University of Szeged.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Róbert Péter: Supervision, Conceptualization, methodology, funding acquisition: Writing – review & editing
Zsolt Szántó: Software, Methodology
Zoltán Biacsi: Software
Gábor Berend: Supervision, Methodology
Vilmos Bilicki: Supervision, Methodology

AUTHOR AFFILIATIONS

Róbert Péter  
Institute of English and American Studies, University of Szeged, Szeged, Hungary
Zsolt Szántó  
MTA-SZTE Research Group on Artificial Intelligence, University of Szeged, Szeged, Hungary
Zoltán Biacsi  
Department of Software Engineering, University of Szeged, Szeged, Hungary
Gábor Berend  
MTA-SZTE Research Group on Artificial Intelligence, University of Szeged, Szeged, Hungary
Vilmos Bilicki  
Department of Software Engineering, University of Szeged, Szeged, Hungary

REFERENCES


TO CITE THIS ARTICLE:

Submitted: 16 October 2023  
Accepted: 18 December 2023  
Published: 07 March 2024

COPYRIGHT:
© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.