Journal of Open Humanities Data

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

DATA PAPER

RÓBERT PÉTER
ZSOLT SZÁNTÓ
ZOLTÁN BIACSI
GÁBOR BEREND
VILMOS BILICKI

*Author affiliations can be found in the back matter of this article

ubiquity press

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:
Péter, R., Szántó, Z., Biacsi, Z., Berend, G., & Bilicki, V. (2024). Multilingual Analysis and Visualization of Bibliographic Metadata and Texts With the AVOBMAT Research Tool. *Journal of Open Humanities Data*, 10: 23, pp. 1–10. DOI: https://doi.org/10.5334/johd.175

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.

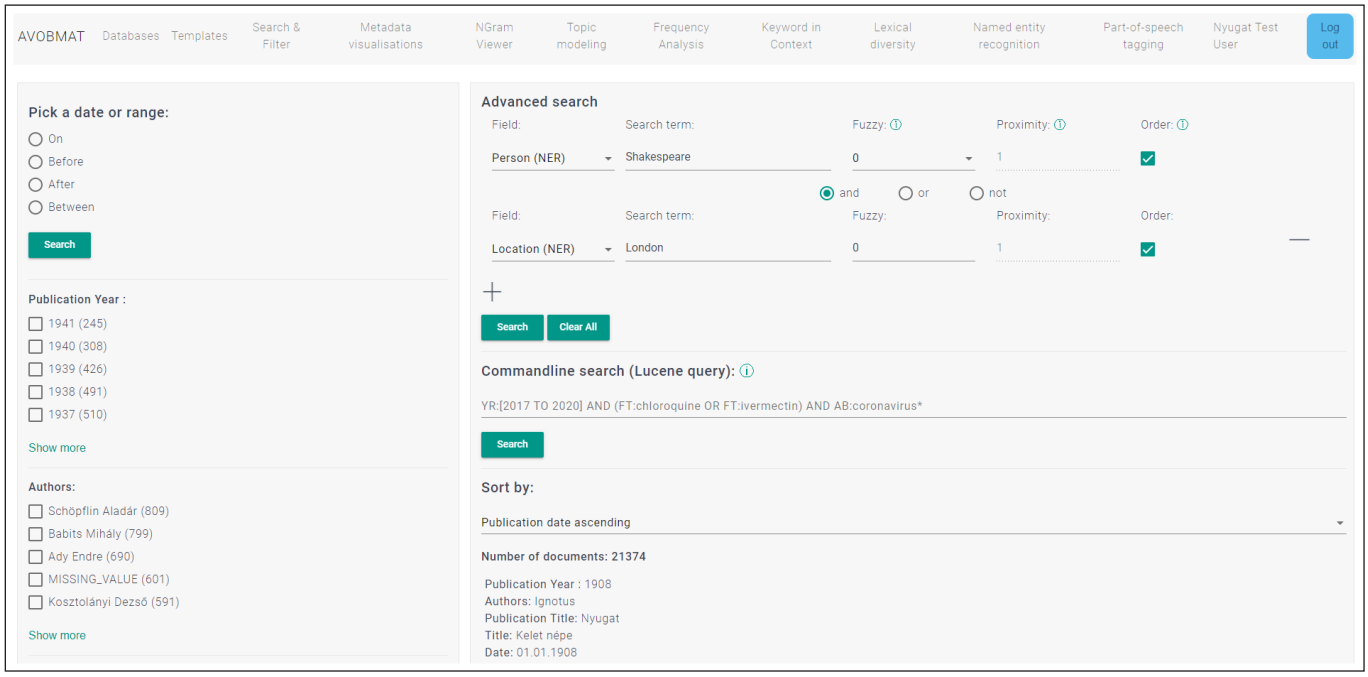


Figure 1 AVOBMAT graphical interface.

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.

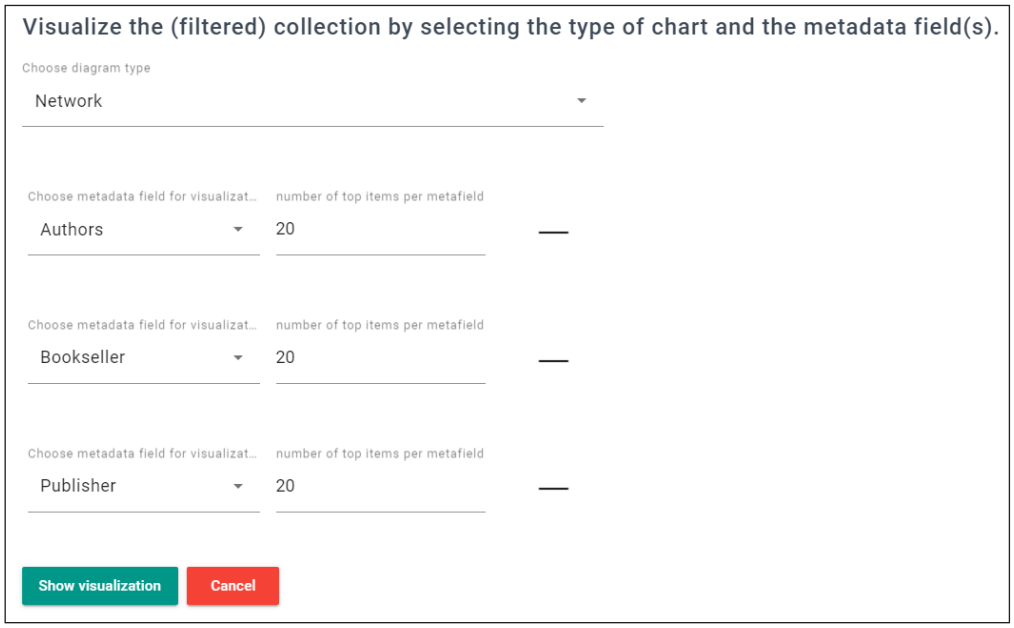


Figure 2 Interactive metadata visualization setting.

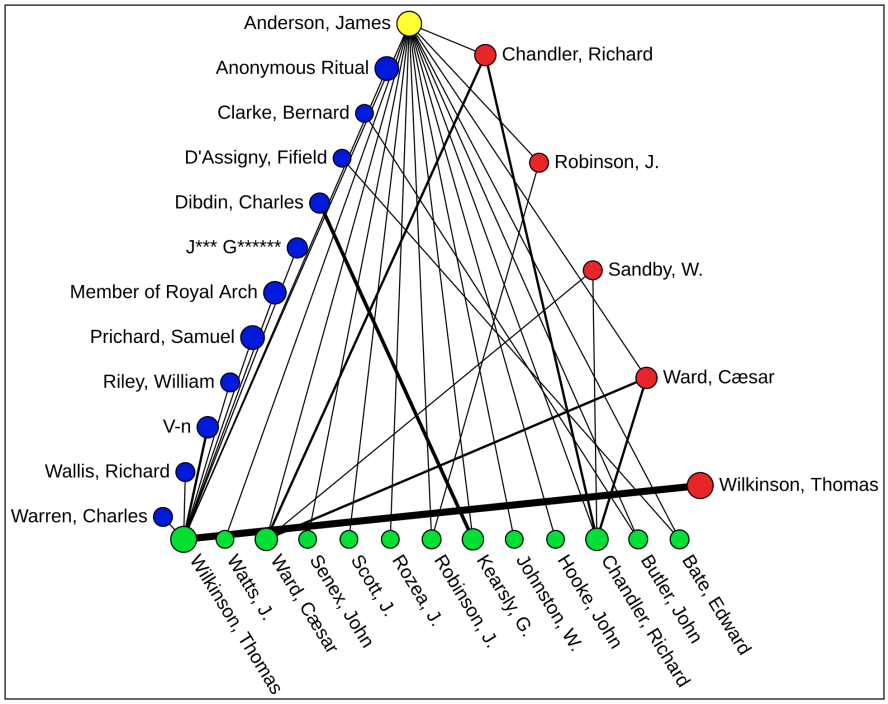


Figure 3 Network analysis of authors, publishers and booksellers involved in the publications of 18th-century books concerning Freemasonry with a particular focus on James Anderson (author).

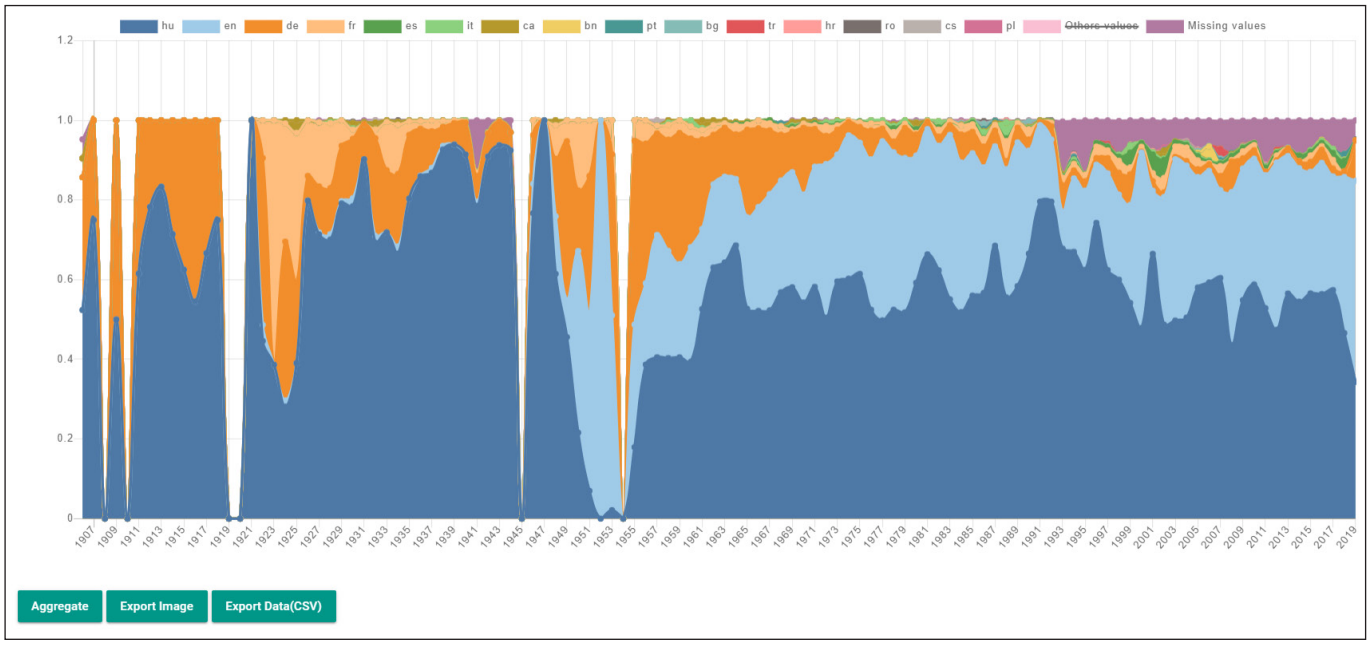


Figure 4 Chronological distribution of the detected languages of the 53411 articles and books in the University of Szeged publication repository.

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.

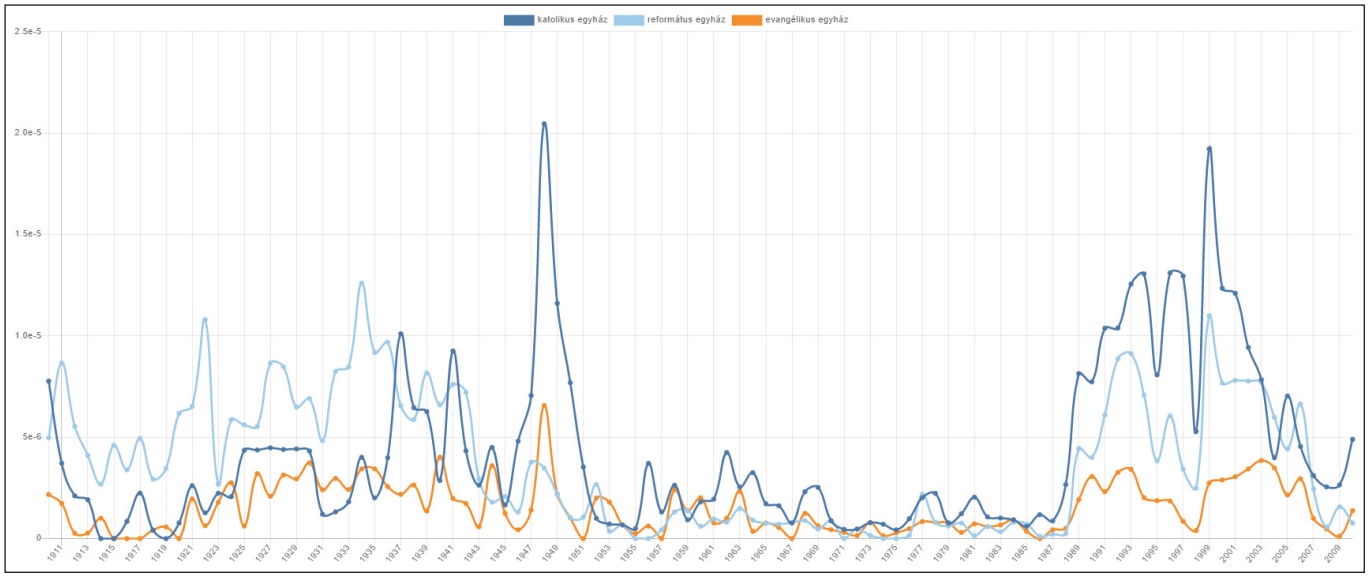


Figure 5 N-gram viewer. Distribution of "katolikus egyház" [Catholic church], "református egyház" [Reformed church], "evangélikus egyház" [Lutheran church] in the *Délmagyarország* daily newspaper, 1911–2009.

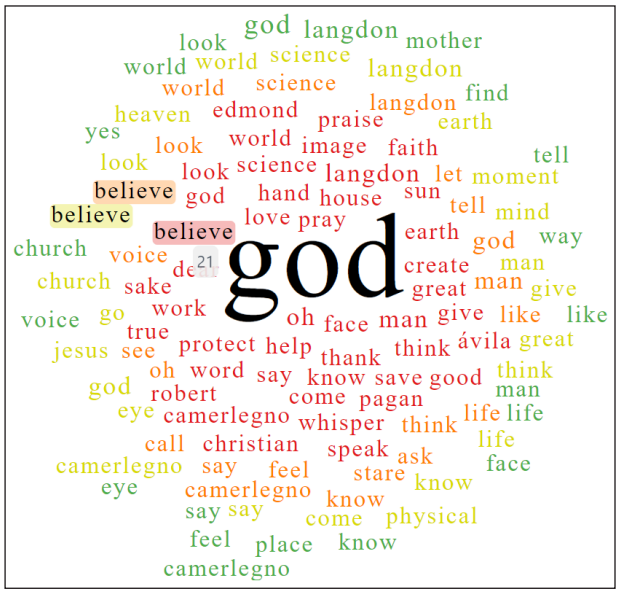


Figure 6 TagSphere analysis. Dan Brown's novels. Keyword: god, word distance: 4 (shown in different colours), minimum word frequency: 7, lemmatized texts, stopwords removed.

Wordcloud

Word	Word distance	Count	%
god	0	924	100.00%
oh	1	32	3.46%
langdon	1	28	3.03%
man	1	27	2.92%
thank	1	22	2.38%
believe	1	21	2.27%
work	1	20	2.16%
hand	1	19	2.06%
think	1	18	1.95%
help	1	17	1.84%
love	1	17	1.84%
create	1	16	1.73%
say	1	15	1.62%
image	1	15	1.62%
sake	1	15	1.62%
god	1	14	1.52%
house	1	13	1.41%
true	1	12	1.30%
pray	1	11	1.19%
know	1	10	1.08%
come	1	10	1.08%
science	1	10	1.08%
face	1	9	0.97%

Export Data

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).

Pu...	Title	Authors	#Toke...	#Types	TTR	Root TTR	Herdan's C	Mass TTR	MSTTR	MATTR	HDD	MTLD
1997	Harry Potter and the Philosopher's Stone	Joanne ...	77591	5672	0.07310126	20.362474	0.76766485	0.047514137	0.82715666	0.82715666	0.8837512	104.93297
1998	Harry Potter and the Chamber of Secrets	Joanne ...	85366	6795	0.079598434	23.256641	0.77711785	0.045197584	0.8330404	0.8330404	0.88406277	109.88616
1999	Harry Potter and the Prisoner of Azkaban	Joanne ...	107504	7336	0.068239324	22.374174	0.7682634	0.046057854	0.82876277	0.82876277	0.8837456	106.605194
2000	Harry Potter and the Goblet of Fire	Joanne ...	191045	10037	0.05253736	22.963394	0.75771654	0.045877147	0.8259005	0.8259005	0.8836171	103.66941
2003	Harry Potter and the Order of the Phoe...	Joanne ...	257217	11897	0.046252776	23.457819	0.7532739	0.04560304	0.83290046	0.83290046	0.88663274	110.80274
2005	Harry Potter and the Half-Blood Prince	Joanne ...	169681	10079	0.059399698	24.468134	0.76552546	0.04483575	0.8324138	0.8324138	0.8841736	108.779274
2007	Harry Potter and the Deathly Hallows	Joanne ...	198613	10874	0.05474969	24.399757	0.7618693	0.04494722	0.823142	0.823142	0.87844294	102.54282

Figure 8 Lexical diversity metrics in J. K. Rowling's Harry Potter novels.

7.4. KEYWORD-IN-CONTEXT

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

View occurances line by line

Authors	Title	Publicatio...	Text
Joanne Rowling	Harry Potter and the Philosopher's Stone	1997	note that the forest on the grounds is forbidden to all pupils. And a few of our older students would do well to remember that as well." Dumbledore's twinkling eyes flashed in the direction of the Weasley twins. "I have also been asked by Mr. Filch, the caretaker, to remind you all that no ***magic*** should be used between classes in the corridors. "Quidditch trials will be held in the second week of the term. Anyone interested in playing for their House teams should contact Madam Hooch. "And finally, I must tell you that this year, the third-floor corridor on the right-hand side is out
Joanne Rowling	Harry Potter and the Philosopher's Stone	1997	Everybody finished the song at different times. At last, only the Weasley twins were left singing along to a very slow funeral march. Dumbledore conducted their last few lines with his wand and when they had finished, he was one of those who clapped loudest. "Ah, music," he said, wiping his eyes. "A ***magic*** beyond all we do here! And now, bedtime. Off you trot!" The Gryffindor first years followed Percy through the chattering crowds, out of the Great Hall, and up the marble staircase. Harry's legs were like lead again, but only because he was so tired and full of food. He was too sleepy even to be
Joanne Rowling	Harry Potter and the Philosopher's Stone	1997	anyone (except perhaps the Weasley twins) and could pop up as suddenly as any of the ghosts. The students all hated him, and it was the dearest ambition of many to give Mrs. Norris a good kick. And then, once you had managed to find them, there were the classes themselves. There was a lot more to ***magic***, as Harry quickly found out, than waving your wand and saying a few funny words. They had to study the night skies through their telescopes every Wednesday at midnight and learn the names of different stars and the movements of the planets. Three times a week they

Export Data

Figure 9 Keyword-in-context. The word "magic" in J. K. Rowling's *Harry Potter and the Philosopher's Stone*.

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.

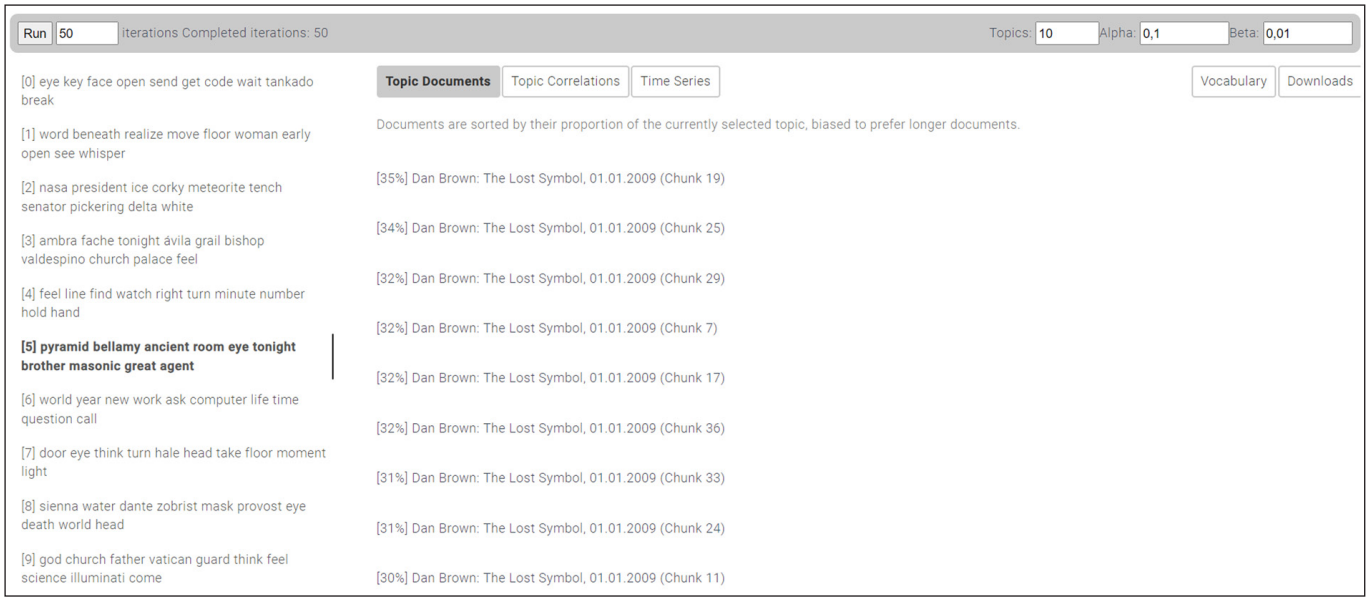


Figure 10 Topic modelling of Dan Brown's novels.

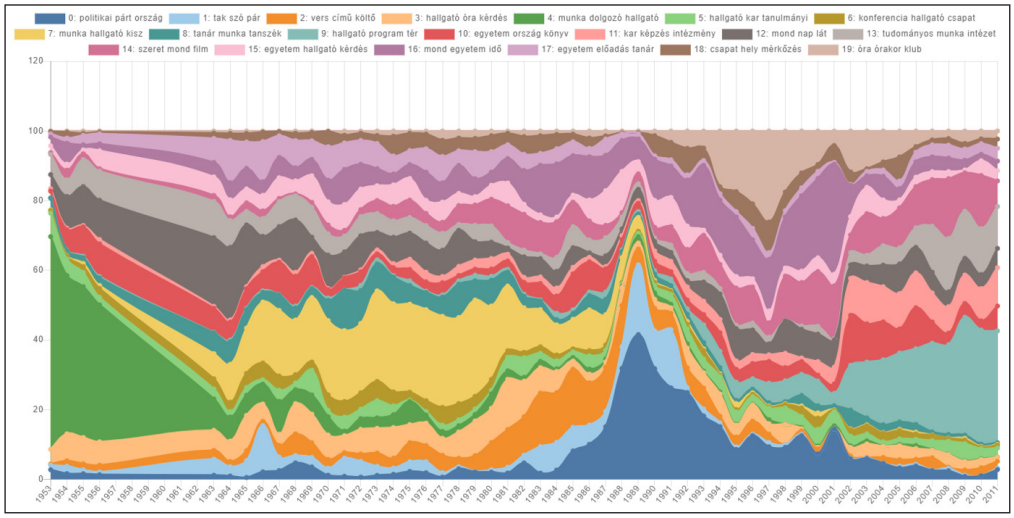


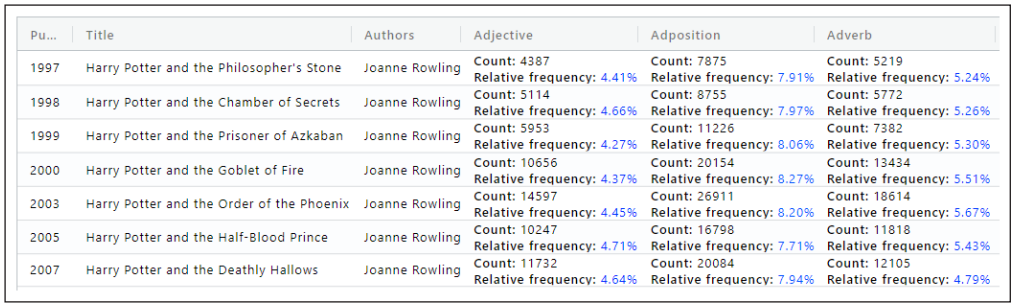
Figure 11 Topic modelling of the *Szegedi Egyetem* [University of Szeged] Magazin, 1953–2011.

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).

Word Form	Lemma	Part-of-speech Tag	Relative Frequency	Count ↓	Documents
symbols	symbol	NOUN	0.001437455	187	7
fire	fire	NOUN	0.001437455	187	7
considered	consider	VERB	0.001437455	187	7
carefully	carefully	ADVERB	0.001437455	187	7
plane	plane	NOUN	0.001429768	186	7
killed	kill	VERB	0.001429768	186	7
bottom	bottom	NOUN	0.001422081	185	7
search	search	NOUN	0.001422081	185	7
simple	simple	ADJECTIVE	0.001422081	185	7
different	different	ADJECTIVE	0.001422081	185	7
things	thing	NOUN	0.001414395	184	7
laughed	laugh	VERB	0.001414395	184	7
leave	leave	VERB	0.001414395	184	7
hung	hang	VERB	0.001414395	184	7
shouted	shout	VERB	0.001414395	184	7

Figure 12 Part-of-speech analysis in Dan Brown's novels.



Pu...	Title	Authors	Adjective	Adposition	Adverb
1997	Harry Potter and the Philosopher's Stone	Joanne Rowling	Count: 4387 Relative frequency: 4.41%	Count: 7875 Relative frequency: 7.91%	Count: 5219 Relative frequency: 5.24%
1998	Harry Potter and the Chamber of Secrets	Joanne Rowling	Count: 5114 Relative frequency: 4.66%	Count: 8755 Relative frequency: 7.97%	Count: 5772 Relative frequency: 5.26%
1999	Harry Potter and the Prisoner of Azkaban	Joanne Rowling	Count: 5953 Relative frequency: 4.27%	Count: 11226 Relative frequency: 8.06%	Count: 7382 Relative frequency: 5.30%
2000	Harry Potter and the Goblet of Fire	Joanne Rowling	Count: 10656 Relative frequency: 4.37%	Count: 20154 Relative frequency: 8.27%	Count: 13434 Relative frequency: 5.51%
2003	Harry Potter and the Order of the Phoenix	Joanne Rowling	Count: 14597 Relative frequency: 4.45%	Count: 26911 Relative frequency: 8.20%	Count: 18614 Relative frequency: 5.67%
2005	Harry Potter and the Half-Blood Prince	Joanne Rowling	Count: 10247 Relative frequency: 4.71%	Count: 16798 Relative frequency: 7.71%	Count: 11818 Relative frequency: 5.43%
2007	Harry Potter and the Deathly Hallows	Joanne Rowling	Count: 11732 Relative frequency: 4.64%	Count: 20084 Relative frequency: 7.94%	Count: 12105 Relative frequency: 4.79%

Figure 13 Part-of-speech analysis of J. K. Rowling's Harry Potter novels. Statistical results.

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).

Language model	Person	Organization	Location	Miscellaneous	Language	Work of art	Geopolitical entity	National or religious group	Date	Ordinal number	Product	Quantity	Time	Money	Infrastructure	Cardinal number	Event	Law	Percent	Period	Movement	Phone	Pet name	Title affix
Chinese	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X					
Danish	X	X	X	X																				
Dutch	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X					
English	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X					
French	X	X	X	X																				
German	X	X	X	X																				
Greek	X	X	X				X				X						X							
Italian	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X
Japanese	X	X	X	X																				
Lithuanian	X	X	X				X				X		X											
Hungarian	X	X	X	X																				
Norwegian Bokmål	X	X	X				X		X				X											
Polish	X	X	X	X																				
Portuguese	X	X	X		X	X	X	X	X	X	X			X	X	X	X			X				
Romanian	X	X	X	X																				
Spanish	X	X	X	X																				
Multi-language	X	X	X	X																				

Table 1 Named entity recognition in different languages.

Figure 14 Named entity recognition and linking in Dan Brown's *Da Vinci Code*.

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-EGYETEMI-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 orcid.org/0000-0002-7972-4751
Institute of English and American Studies, University of Szeged, Szeged, Hungary

Zsolt Szántó orcid.org/0000-0002-8924-206X
MTA-SZTE Research Group on Artificial Intelligence, University of Szeged, Szeged, Hungary

Zoltán Biacsi orcid.org/0009-0000-7204-2865
Department of Software Engineering, University of Szeged, Szeged, Hungary

Gábor Berend orcid.org/0000-0002-3845-4978
MTA-SZTE Research Group on Artificial Intelligence, University of Szeged, Szeged, Hungary

Vilmos Bilicki orcid.org/0000-0002-7793-2661
Department of Software Engineering, University of Szeged, Szeged, Hungary

REFERENCES

Blei, D. M., Ng, Y. A., & **Jordan, M. I.** (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Covington, M. A., & **McFall, J. D.** (2010). Cutting the Gordian Knot: the Moving-Average Type-Token Ratio (MATTR). *Journal of Quantitative Linguistics, 17*(2), 94–100. DOI: https://doi.org/10.1080/09296171003643098

Jänicke, S., & **Scheuermann, G.** (2017). On the Visualization of Hierarchical Relations and Tree Structures with TagSpheres. In: Braz, J, et al. (Eds.), *Computer Vision, Imaging and Computer Graphics Theory and Applications*. Cham: Springer International Publishing. pp. 199–219. DOI: https://doi.org/10.1007/978-3-319-64870-5_10

Juršic, M, et al. (2010). Lemmagen: Multilingual Lemmatisation with Induced Ripple-down Rules. *Journal of Universal Computer Science, 16*(9), 1190–1214.

Manning, C. D., Raghavan, P., & **Schütze, H.** (2009). *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511809071

McCarthy, P. M., & **Jarvis, S.** (2010). MTLD, vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment. *Behaviour Research Methods, 42*(2), 381–392. DOI: https://doi.org/10.3758/BRM.42.2.381

Péter, R., Szántó, Zs., Seres, J., Bilicki, V., & **Berend, G.** (2020). AVOBMAT: a digital toolkit for analysing and visualizing bibliographic metadata and texts. In: G. Berend, G. Gosztolya & V. Vincze (Eds.), *XVI. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem, Informatikai Intézet, pp. 43–55.

Péter, R., Szántó, Zs., Seres, J., Bilicki, V., & **Berend, G.** (2022). Az AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) többnyelvű kutatási eszköz bemutatása. *Digitális Bölcsészet*, 4, 3–28. DOI: https://doi.org/10.31400/dh-hun.2021.4.3530

Rudi, L. C., & **Vitányi, P. M. B.** (2007). The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering, 19*(3), 370–383, https://arxiv.org/pdf/cs/0412098v3.pdf. DOI: https://doi.org/10.1109/TKDE.2007.48

Significant text aggregation. Available at https://www.elastic.co/guide/en/elasticsearch/reference/8.0/search-aggregations-bucket-significanttext-aggregation.html [Last accessed 13 October 2023].

SpaCy Models and Languages. Available at https://spacy.io/usage/models [Last accessed 13 October 2023].

Torruella, J., & **Capsada, R.** (2013). Lexical Statistics and Tipological Structures: A Measure of Lexical Richness. *Procedia: Social and Behavioral Sciences, 95*, 447–454. DOI: https://doi.org/10.1016/j.sbspro.2013.10.668

TO CITE THIS ARTICLE:
Péter, R., Szántó, Z., Biacsi, Z., Berend, G., & Bilicki, V. (2024). Multilingual Analysis and Visualization of Bibliographic Metadata and Texts With the AVOBMAT Research Tool. *Journal of Open Humanities Data*, 10: 23, pp. 1–10. DOI: https://doi.org/10.5334/johd.175

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

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