Use the Metadata, Luke! – An Experimental Joint Metadata Search and N-gram Trend Viewer for "Personal" Web Archives

Balázs Indig, Zsófia Sárközi-Lindner, Mihály Nagy
Eötvös Loránd University, Department of Digital Humanities
National laboratory for Digital Humanities
Muzeum krt. 6-8., H-1088, Budapest, Hungary
{indig.balazs,lindner.zsofia,nagy.mihaly}@btk.elte.hu

Abstract

Many digital humanists (philologists, historians, sociologists, librarians, the audience for web archives) design their research around metadata (publication date ranges, sources, authors, etc.). However, current major web archives are limited to technical metadata while lacking high quality, descriptive metadata allowing for faceted queries. As researchers often lack the technical skill necessary to enrich existing web archives with descriptive metadata, they increasingly turn to creating personal web archives that contain such metadata, tailored to their research requirements. Software that enable creating such archives without advanced technical skills have gained popularity, however, tools for examination and querying are currently the missing link. We showcase a solution designed to fill this gap.

1 Introduction

The potential of the vast amount of data generated on the World Wide Web (as a corpus of text) has taken the lead from printed media and is constantly evolving (e.g. the U.S. presidential tweets). For example, sociologists and linguists must use online sources to gain insight into recent trends (temporal, ideological standpoints) in social topics and language as the print media can not keep up the pace and provide real-time updates (e.g. COVID-19 news, social media). Luckily, the state of major web archives indicates that all important data will likely be preserved. Therefore, beyond tackling the difficulties of web harvesting, meeting the needs of researchers has become the most recent goal of archivists (Major and Gomes, 2021, 13).

However, the quantity of data in major archives does not counteract deficiencies and negligence in quality. The aforementioned direction of research requires quality descriptive metadata (publication date, author, column, portal name, keywords, etc.) for specific portals with complete and up-to-date archives, to for example assess source dependent standpoints. Unfortunately, major web archives are limited to full-text or URL searching, even more advanced ones only include filters on technical metadata (e.g. file format, crawl date), and a rare few such as `https://webarchives.ca/` have facets on descriptive metadata.

As Hale et al. (2017), Indig et al. (2020) and Costa (2021) have pointed out, for a more detailed overview of sources, a uniform format for metadata and text would be needed because websites are typically neither consistent nor well-structured. Nevertheless, the complexity and time-consuming nature of isolating the useful text, identifying the segments along with metadata (e.g. title, headings, image), preprocessing (e.g. tokenization and parsing) and postprocessing (e.g. filtering, deduplication, curation of metadata) in existing web archives make it an unattainable promise of future AI solutions. This is especially relevant for those without technical skills or sufficient infrastructure[1].

Fortunately, many workflows (e.g. crawling, text extraction, NLP, plotting) can be conducted sufficiently well with a few clicks on free, easy-to-use software or services that work on uploaded data or prepared data sets (see examples in Section 3.). Researchers with insufficient funds (e.g. PhD students) are therefore limited to creating "personal" web archives using such tools to gian at least preliminary results through exploratory analysis and test their ideas before going further. These attempts usually get stuck at the aggregation and visualisation stages, as available tools and tutorials do not facilitate the freedom to experiment, instead only showcasing basic, and typical workflows. With our tool these scholars can find answers to detailed questions involving metadata based on their gathered data with at most a few line of code needed

[1]Downstream projects e.g. *the OSCAR corpus* (Ortiz Suárez et al., 2019) discard metadata in favour of deduplication and filtering.

The 2nd International Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 47–52
November 20, 2022.

for the collecting and preprocessing steps.

Such archives must be created manually and in many cases cannot be published or shared due to conservative legal regulations. This puts pressure on website operators, as everyone is forced to create an archive of their own, the very issue that was intended to be avoided by shared archives such as *Common Crawl* that have become trapped in a rivalry of quantity over quality. On top of that, the *Internet Archive* does not support downloading and restoring pages[2], and as of yet does not contain a faceted query and visualisation interface for descriptive metadata. In the future, large web archives may catch up and support querying quality uniform descriptive metadata out of the box, but until then scholars must individually experiment with the available tools on a smaller scale. Here we find the eternal struggle to strike a balance between precision and recall in a non-standardised open domain.

2 Background

Ruest et al. (2020) developed *The Archives Unleashed Toolkit*[3] which generalised typical scholarly activities into the *Filter-Extract-Aggregate-Visualize (FEAV) cycle*. This model features a chain of independent tasks where each task can be adapted separately to the technological developments and different goals on different scales. Nonetheless, existing tools are not yet sufficiently easy-to-use and integrated for individual researchers, who opt for simpler ones which are easier to setup, use and adapt.

Off-the-shelf services are too dependent on the provider. Moreover, they are constantly growing and developed, which means that these services will inevitably change over time[4] (e.g. shut down or being extended), and become volatile[5]. This hinders the reproducibility of obtained results, forcing researchers not to trust such sources, since it is essential to ensure the citation and long-term preservation of research data and final research results (Barats et al., 2020). In the long term, research material should be suitable for the widest possible research community, and should stand the test of time to facilitate reproducibility. This requires web archivists to resist the siren voices of merely increasing size, and strive to democratise and further reduce the complexity of the tools involved.

3 Tools for a Modular FEAV Cycle

Tools such as *Trafilatura* (Barbaresi, 2021) serve as good examples, as they are adequate for creating custom, "personal" web archives and have unique features compared to existing major web archives (e.g. extracting descriptive metadata and text in a uniform format with sufficient quality and quantity). In a few lines of Python code one can create enough research data for preliminary analysis, and by combining such data sets more complicated advanced research questions can be answered in depth.

Processing such text with NLP pipelines (e.g. *SpaCy* (Montani et al., 2022)) with another few lines of code to extract lemmas and named entities is easy even for non-technical researchers. On the visualisation front there exist a large number of tools and libraries providing quick and elegant solutions (e.g. *D3.js* (Bostock, 2012)), often implicitly handling most parameters, however, their functionality either focuses on ease-of-use or modularity, rarely both. Their popularity – judging by their download and citation counts – shows that people tend to prefer them. They may also appear as building blocks of some full-fledged compact applications as well. The last remaining part is the aggregation of data, which is either done with predefined workflows limiting its scope or is completely left to the user with all of its complexity.

Following from the design of such pipelines, if any of these tools are not yielding satisfactory results, they can be changed independently (maintaining the FEAV cycle) as their output is in simple and standard format. We take these tools as granted, and focus on the missing part, the ground-up design of an ideal trend viewer as the aggregation module for the desired tool chain.

4 Trend Viewer Trends

The history of n-gram trend viewers starts with the *Google Books N-gram Viewer* (Michel et al., 2011) which has an exceptionally large corpus, but its software is not freely available to try on custom

[2] https://wiki.archiveteam.org/index.php?title=Restoring

[3] https://github.com/archivesunleashed/aut

[4] For example https://archivesunleashed.org/cloud/ has "shut down on 30 June 2021, and will be replaced by a similar service ".

[5] For example *Warcbase*, https://github.com/lintool/warcbase (Lin et al., 2017) is unmaintained since September 2017 in favour of *The Archives Unleashed Toolkit*.)

data. Another similar example is the *Microsoft Web N-gram Services*, which has been shut down[6].

The National Library of Norway also established an elaborate n-gram trend search service (Birkenes et al., 2015) on their collection of books, but the source code[7] has not been updated for 7 years and is written in the now deprecated *Python 2* language. Its *SQLite* database back end can handle 34 billion words from books and newspapers in a remarkable way. The project is still relevant as the *Icelandic Gigaword Corpus* utilises the code in their recent work (Steingrímsson et al., 2020).

The most versatile existing solution is *Shine*[8] (Jackson et al., 2016), an experimental faceted search engine and trend viewer tool based on *Solr* in its early stages, however, it has not shown signs of development since 2020. It has similar goals to those of our project (see Section 1.), including faceted search over various metadata, however, it fails to cover descriptive metadata other than publication date when it comes to aggregation and visualisation. The *Web Archives for Longitudinal Knowledge (WALK)*[9] project is a good demonstration of applying Shine to facilitate DH research.

These examples suggests that there is an emerging trend in the DH and NLP of using descriptive metadata for diachronic corpora (created from e.g. web archives) where available. Such data augmented with modern NLP (e.g. sentiment analysis, wikification (Brank et al., 2017)), aggregated and displayed in the context of not just time, but all other accessible factors can help digital humanists focus and gain insight on new previously nonexistent scopes of their research. Our aggregation module works independently on the output of existing crawling and NLP tools, and uses an existing visualisation tool, therefore maintaining ease of use and sufficient generality.

5 Method: Redesign from the Ground Up

We encourage future researchers to utilise e.g. Trafilatura to create their own research data in standard format (e.g. *TEI XMLs*) that can be lemmatised effectively with e.g. SpaCy to get preprocessed data in decent quality and quantity in almost any language[10]. We also provide example scripts, however, for the sake of reproducibility we used an existing data set containing descriptive metadata for our experiments: Indig et al. (2020) developed a pilot web archive described as a *complete*, *reliable* and *versatile* representation of archived content (recorded in *WARC* files) for Hungarian news portals while *maintaining archive content authenticity* (Lendák et al., 2022). It contains around 1 billion words from about 2.8 million news articles written in the last 20 years composed of roughly 20 different Hungarian news portals with a standardised layout and descriptive metadata in *TEI XML* format. The whole archive is available with DOI links on the *Zenodo.org* repository[11] which is maintained by *CERN* ensuring the reproducibility of further downstream research and making a positive example of legally sharing "personal" web archives for non-profit research purposes.

Building on the various types of available metadata in the input data set (publication date, author, column, portal name, keywords, etc.), we designed an extensible hybrid query service. On one hand, the retrieved records can be viewed as a list of links to the original pages as in traditional search engines, and the query can be refined with filters on metadata fields. Additionally authorised non-profit researchers can view – legally, without copyright infringement – the full article in our standardised HTML format which contains the original text formatting generated on-the-fly from the stored *TEI XMLs*. This allows access even if the original link is broken or its content has changed, which also helps manual work and gathering examples. On the other hand, the features of the retrieved documents can be visualised with different kinds of graphs (e.g. line and bar charts) by setting the appropriate parameters (e.g. the fields for the axes), allowing the user to create custom views of the data filtered by the represented features (see Figure 1.) opening possibilities for non-technical users to express research questions (see Section 7. for more examples).

All retrieved data and plotted graphs can be downloaded for use e.g. in research papers or additional processing. As an example, bundling such a tool with a research paper can enable readers to cross-check and further examine the presented results and enhance reproducibility and reusability,

[6] https://www.microsoft.com/en-us/research/project/web-n-gram-services/

[7] https://github.com/NationalLibraryOfNorway/NB-N-gram

[8] https://github.com/ukwa/shine

[9] https://webarchives.ca/

[10] see the evaluation in the respective papers

[11] https://zenodo.org/communities/elte-dh/search?page=1&size=20&q=TEI

increasing the impact of the paper.

6 Technical Evaluation and Details

Existing search engine platforms e.g. *Solr* or *ElasticSearch* provide a solid foundation, however, their complexity and possible vulnerabilities are prone to hacking attempts and require continuous maintenance. Their complexity also discourages external developers from customisation attempts. The same holds for deep learning frameworks, which also require advanced technical skills and machinery.

Birkenes et al. (2015) have shown that storing n-grams in *SQLite* can be fast enough even for databases extending over 34 billion tokens and yield correct results. However, they use separate tables for uni-, bi- and trigrams. Our approach was to add padding tokens at the end of text segments and create only a table with 5-grams, as we can represent lower-order n-grams using only the first n tokens, thus creating fewer rows overall. The actual frequencies come from grouping the matching lower-order n-grams (*GROUP BY* statement in SQL) and *SUM*ming the frequencies for each group[12]. So one can easily use wider search patterns e.g. "[first name] [surname] [job title] [organisation]".

Each n-gram has a frequency and is linked to a document while the metadata (publication date, sources, authors, etc.) for each document – identified by a document ID – are stored in a row in a separate table along with the path for the TEI XML document on the file system. We distinguish three types of fields: a) simple text (e.g. portal name, column), b) multi-valued text (e.g. author, keywords), c) date (e.g. publish date) which can specified as an interval when used as a filter.

Our import script requires a *CoNLL-U* like vertical TSV format used by Sketch Engine (Kilgarriff et al., 2014). The possible lemmas for each word are acquired by lemmatising the text prior to database creation with a common lemmatiser. We only store the lemma n-grams in the database, because their disambiguation in the query string is done by a look-up table using all possibilities in *conjunctive normal form (CNF)* as ambiguous n-grams are unlikely and results close to real lemmatisation can be achieved.

Our pipeline uses *AWK* and standard *Unix* tools, allowing to create and count lemma 5-grams for 1 billion words while retaining their source document IDs, in about an hour. The user has to write a configuration file and run a single command to import the data into the database. We provide example scripts to help users from the crawling through to the importing stages. To allow rewriting or different development paths, we decoupled the front end and the back end which communicate with a stable REST API. The application is available as a docker image to facilitate operating system independence and long term usage.

The only limiting factor we experienced is the quality of the used data, as regarding performance even low-end hardware can handle databases with the the expected size.

7 Examples

Search engines for web archives or interfaces for visualising trends usually help researchers navigate through data sets, allowing quick assessment of the usefulness of materials for current research (e.g. cross-check data to reveal artefacts), possibly verifying simpler claims or filtering useful parts (saving time by not processing unused segments). We would like to illustrate how, thanks to more detailed preprocessing, a few tiny adjustments can expand the possibilities for answering more complex research questions. We provide presets in the WebUI for users to show the examples given below.

One can observe the emergence and rivalry of new terms across different news portals, e.g. in relation to the climate crisis[13], or the way how medical jargon[14] became common language during the COVID-19 pandemic (Varga et al., 2022a,b), and how the correlation between its key issues (e.g. *restriction*, *quarantine* vs. *vaccination*) have changed over time by the start of the third wave. Another example demonstrates how one can get a quick answer to the question "Which values (*national*, *chatolic*, *european*) are more prominent in the articles of the selected news portals?", which can then be nuanced through more detailed examination (e.g. their ideological standpoints). A small but important detail is that distortions due to the size of each portal can be mitigated by expressing the occurrences of n-grams as a percentage (e.g. percentage

[12] A prefix trie or perfect hash based approaches could potentially yield more compact storage format, but they would increase complexity as metadata still would require a database.

[13] e.g. *climate change*, *climate emergency*, *climate strike*, *climate hysteria*

[14] e.g. *social distancing*, *spike protein*, *long covid*, *PCR test*, *super-spreader*

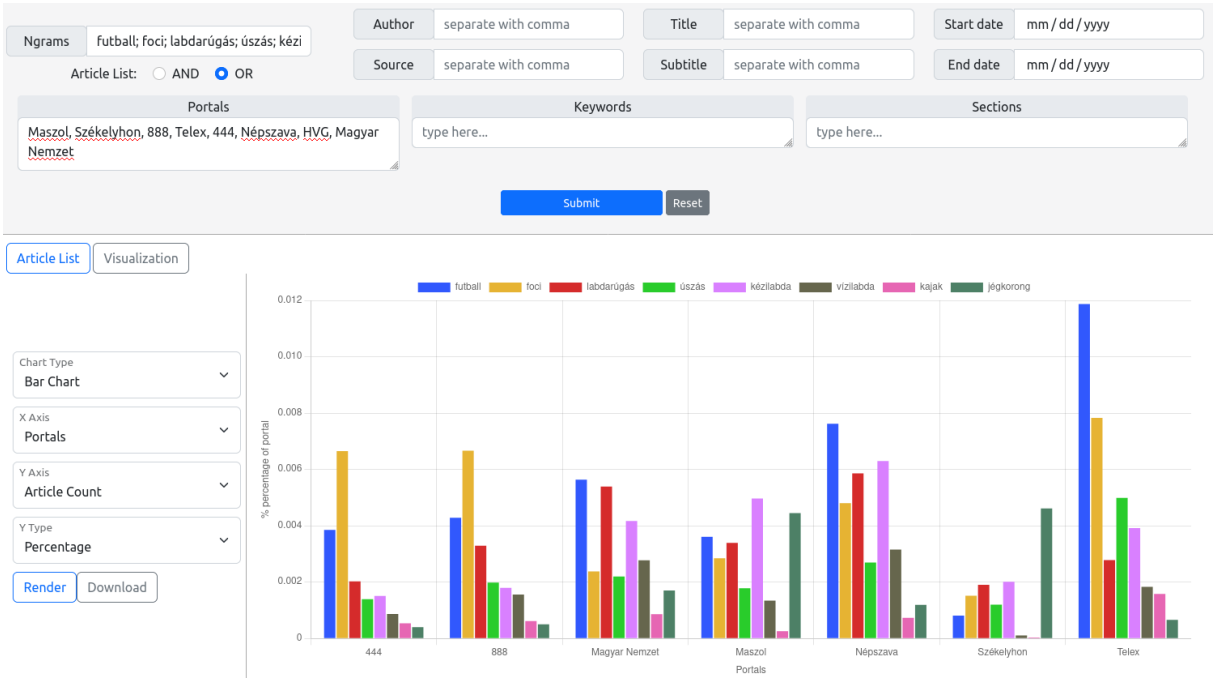

Figure 1: The distribution of mentions of different kinds of sport across portals on the WebUI.

of articles per portal). When listing relevant articles the logical operators applied to the n-gram can be used to separate the trend of co-occurrence and independent occurrences of the listed n-grams. Thanks to accurate metadata extraction, the use of hashtags by online newspapers and journalists can be tested, the variety and consistency of keywords and phrases associated with topics can be a telling feature, and it helps to improve the accuracy of the search and the weighting of the topic. In the field of sports, we can investigate the interest of some Hungarian news portals in Hungary and abroad, by querying articles about different sports (see Figure 1). We would expect that portals in Hungary would be predominantly concerned with football results and in Romania *hockey* (*jégkorong*) would be more prominent (e.g. on *maszol.ro*). The fact that in Hungarian three synonyms of almost equal rank are typically used to describe football (*labdarúgás*, *futball*, *foci*) makes interpretation difficult, because synonyms sometimes complement each other and sometimes overlap, which is reflected in the number of results. This can be clarified with the downloadable TSV files. Future plans include making it even easier to do this directly in the search engine by utilising vector space models to unify synonyms if enabled. A further point to note is the importance of context. If researchers want to quantify traces of hate speech or ethnic exclusion, they need to be familiar with the background and the attitudes of the sources used. A simple but striking example is the occurrence of the term *gypsy* (*cigány*). In two of the newspapers studied, the results are very prominent, but one is an avowedly extremist medium (far-right wing), while the other deals specifically with social issues, education and inclusion.

8 Conclusion and Discussion

We introduced a search and visualisation engine for viewing n-gram trends in light of different descriptive metadata implemented in a lightweight SQL-based framework. A custom data set can be imported with a few lines of Python code, and the customised visualisation can be exported to be used directly in research papers. This helps researchers express their questions freely or share their data through a query and visualisation interface (e.g. as anonymous supplementary material for research papers) and enable further examination and possible new discoveries on the same data set.

We plan to add more NLP modules (e.g. sentiment analysis to study hate speech and wikification for semantic results) for different layers of annotation and support for CQL expressions. The source code[15] and docker image is published under a copyleft license and a public pilot service is available.

[15] https://github.com/ELTE-DH/meta-trend-viewer

References

Christine Barats, Valérie Schafer, and Andreas Fickers. 2020. Fading Away... The challenge of sustainability in digital studies. *Digital Humanities Quarterly*, 014(3).

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.

Magnus Breder Birkenes, Lars G. Johnsen, Arne Martinus Lindstad, and Johanne Ostad. 2015. From digital library to n-grams: NB n-gram. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 293–295, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Mike Bostock. 2012. D3.js - data-driven documents.

Janez Brank, Gregor Leban, and Marko Grobelnik. 2017. Annotating documents with relevant wikipedia concepts. In *Proceedings of the Slovenian Conference on Data Mining and Data Warehouses (SiKDD 2017)*, Ljubljana, Slovenia.

Miguel Costa. 2021. *Full-Text and URL Search Over Web Archives*, pages 71–84. Springer International Publishing, Cham.

Scott A. Hale, Grant Blank, and Victoria D. Alexander. 2017. *Live versus archive: Comparing a web archive to a population of web pages*, pages 45–61. UCL Press.

Balázs Indig, Árpád Knap, Zsófia Sárközi-Lindner, Mária Timári, and Gábor Palkó. 2020. The ELTE.DH pilot corpus – creating a handcrafted Gigaword web corpus with metadata. In *Proceedings of the 12th Web as Corpus Workshop*, pages 33–41, Marseille, France. European Language Resources Association.

Andrew Jackson, Jimmy Lin, Ian Milligan, and Nick Ruest. 2016. Desiderata for exploratory search interfaces to web archives in support of scholarly activities. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, JCDL '16, page 103–106, New York, NY, USA. Association for Computing Machinery.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, pages 7–36.

Imre Lendák, Balázs Indig, and Gábor Palkó. 2022. WARChain: Consensus-based trust in web archives via proof-of-stake blockchain technology. *Journal of Computer Security*, pages 1–17.

Jimmy Lin, Ian Milligan, Jeremy Wiebe, and Alice Zhou. 2017. Warcbase: Scalable analytics infrastructure for exploring web archives. *Journal on Computing and Cultural Heritage*, 10(4).

Daniela Major and Daniel Gomes. 2021. *Web Archives Preserve Our Digital Collective Memory*, pages 11–19. Springer International Publishing, Cham.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O'Leary McCann, Maxim Samsonov, Jim Geovedi, Jim O'Regan, Duygu Altinok, György Orosz, Søren Lind Kristiansen, Daniël de Kok, Lj Miranda, Roman, Explosion Bot, Leander Fiedler, Grégory Howard, Edward, Wannaphong Phatthiyaphaibun, Richard Hudson, Yohei Tamura, Sam Bozek, murat, Ryn Daniels, Peter Baumgartner, Mark Amery, and Björn Böing. 2022. explosion/spaCy: New Span Ruler component, JSON (de)serialization of Doc, span analyzer and more.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora (CMLC-7)*, Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Nick Ruest, Jimmy Lin, Ian Milligan, and Samantha Fritz. 2020. *The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives*, Joint Conference on Digital Libraries 2020 (JCDL 2020), page 157–166. Association for Computing Machinery, New York, NY, USA.

Steinþór Steingrímsson, Starkaður Barkarson, and Gunnar Thor Örnólfsson. 2020. Facilitating corpus usage: Making Icelandic corpora more accessible for researchers and language users. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3399–3405, Marseille, France. European Language Resources Association.

Éva Katalin Varga, Emese Márton, Balázs Indig, Zsófia Sárközi-Lindner, and Gábor Palkó. 2022a. Erdélyi és anyaországi orvosi terminológia pandémia idején. *Alkalmazott Nyelvtudomány*, page in press.

Éva Katalin Varga, Ákos Zimonyi, Balázs Indig, Zsófia Sárközi-Lindner, and Gábor Palkó. 2022b. Durva influenza vagy veszélyes világjárvány? a covid-19 terminológiája a médiában. *XXVIII. Magyar Alkalmazott Nyelvészeti Kongresszus : Nyelvek, Nyelvváltozatok, Következmények*, page in press.