Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1817–1823
Marseille, 20-25 June 2022

Annotating Attribution in Czech News Server Articles

Barbora Hladká[1], Jiří Mírovský[1], Matyáš Kopp[1], Václav Moravec[2]
[1]Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
[2]Charles University, Faculty of Social Sciences, Institute of Communication Studies and Journalism
{hladka, mirovsky, kopp, }@ufal.mff.cuni.cz, vaclav.moravec@fsv.cuni.cz

Abstract

This paper focuses on detection of sources in the Czech articles published on a news server of Czech public radio. In particular, we search for attribution in sentences and we recognize attributed sources and their sentence context (signals). We organized a crowdsourcing annotation task that resulted in a data set of 2,167 stories with manually recognized signals and sources. In addition, the sources were classified into the classes of named and unnamed sources.

Keywords: media bias, news server, article, attribution, signal, source, annotation, crowdsourcing

1. Introduction

Automated journalism refers to the use of algorithms to automatically generate news from structured data, see e.g., Leppänen et al. (2017). Artificial Intelligence (AI) journalism is a broader concept that includes not only automation but machine learning and data processing of various newsrooms related tasks as well. In his survey, Becket (2019) reports that news gathering, news production, and news distribution are the three most common areas for his respondents' future AI-tool wishlist. Marconi (2020) provides readers with a more detailed discussion on how journalism will change through the AI-based processes. No doubt Natural Language Processing (NLP) plays (and will play) a key role in AI journalism.

One of the areas that is currently receiving a lot of attention is the area of a systematic, empirically-based, and historical-comparative understanding of media bias. Wikipedia defines media bias as "the perceived bias of journalists and news producers within mass media in the selection of events and stories that are reported, and how they are covered."[1] Quite surprisingly, there are still only few NLP studies systematically analyzing media bias, see e.g., Hamborg et al. (2019). Currently, significant attention is being paid to fact-checking, see e.g., Thorne and Vlachos (2018), Lazarski et al. (2021). In social sciences, the news production process is an established model that defines nine forms of media bias and describes where these forms originate from, see e.g., Baker (1996), Park et al. (2009).

An informative, balanced article should provide the background of a story, including naming sources. This paper aims at a news server of Czech public radio and focuses on detection of sources in its articles. In particular, we search for attribution in sentences and we extract attributed sources from them. This is a task from the area of text understanding and, at least to our knowledge, an NLP system automatically processing attribution in newspapers articles has not been implemented yet. Prasad et al. (2006) have proposed and described an annotation scheme for marking the attribution in the Penn Discourse TreeBank. However, we plan to use morphological and syntactic relations to detect attributed sources.

Attribution The Macmillan Dictionary defines *attribution* as "the act of attributing something to a particular cause or person, especially the act of saying that something was written, said, painted etc. by a particular person".[2]

Our primary task is to automatically detect sources that journalists credit in newspaper stories. We will approach it by detecting a sentence context in which sources are attributed and therefore we formalize the definition of *attribution* as follows

$$attribution = source + information + signal$$

where *source* originally provided *information* and *signal* is a textual marker that identifies the *source* of the *information*. We use mathematical notation intentionally, namely to emphasize that the order of *source*, *information* and *signal* in the sentence does not play a role which is analogous to the commutative property of addition. For illustration, we recognize the information *there are 7.77 million Internet users over the age of ten in the Czech Republic* and the source *Netmonitor* attributed using the signal *according to* in the sentence *According to NetMonitor, there are 7.77 million Internet users over the age of ten in the Czech Republic.*

In English grammar, a *signal phrase* is a phrase, clause, or sentence that introduces a quotation, paraphrase, or summary, e.g., *Marianne Egeland, Professor of Comparative Literature at the University of Oslo, argues that*. We found it in several teaching materials that clearly explain their motivation to use the word *signal*:

[1] https://en.wikipedia.org/wiki/Media_bias

[2] https://www.macmillandictionary.com/dictionary/british/attribution

"*They* [signal phrases] *signal to a reader that the writer is using an outside source*".[3] Signal phrases inspired us and we modified the *signal* definition so that we consider *signal* and *source* separately.

The rest of the paper is organized as follows: A classification hierarchy of sources is presented in Section 2. In Section 3, we describe an automatic pipeline for processing the iRozhlas collection of stories published by the news server of Czech public radio. A subset of this collection, the SIR 1.0 corpus, was annotated in the annotation task described in Section 4 and evaluated in Section 5. We conclude by summarizing future plans to analyse the annotated data in great detail and to generate signal patterns to detect sources in the complete iRozhlas collection.

2. Source Classification

Named sources Their attribution is as descriptive as possible. They can be further divided according to affiliated institutions into:

- **Official sources** One of their main characteristics is their authority and importance hierarchy. These sources not only have access to information, but make political, economic and social decisions as well. They usually have a dominant position among journalistic sources since both journalists and news consumers attribute higher information quality to them, which need not always be a legitimate expectation. Typically their positions and institutions are mentioned. They can be further classified as:
  - **Political sources** include political actors according to their political party affiliation . We can also include politicians representing executive and legislative bodies (i.e. president, prime minister, ministers, senators, deputies, etc.), e.g., *member of Parliament Jaroslav Faltýnek/ANO/*, *ODS chairman Petr Fiala*.
  - **Non-political sources** include sources usually connected to specific institutions and positions, e.g., *director of the war museum*.
- **Unofficial sources** do not have as much authority as official sources. They are often "ordinary people" as witnesses of important events or confidential information providers (e.g., *experts*, *most scientists*). Unofficial sources are essential for the development of investigative journalism, which often guarantees their anonymity and confidentiality. In this case, as compared with official sources, editorial routines are associated with a more careful verification of information.

Unnamed sources The name (and surname) of a source, their occupation, their affiliated institution, etc. are not mentioned in the text. Journalists usually guarantee their anonymity because of their security. In journalism, two degrees of anonymity are generally distinguished:

- **Completely anonymous sources** are typically attributed using the phrases *unnamed/reliable/anonymous/secret source(s)*
- **Partially anonymous sources** are typically attributed using the phrases *a source close to* . . . , *a source familiar with* . . .

3. iRozhlas Collection

iRozhlas is a news server of the Czech public radio launched on April 18, 2017.[4] The iRozhlas collection where we will detect sources contains 63,325 articles from the period April 18,2017–June 24,2021 and written by 927 authors. All the articles are Czech.

Originally, the iRozhlas collection was represented in the JSON format containing the following items for each story: Story identifier, URL, Date of publication and change, Domicile, List of authors, List of sections, List of tags, Headline, Leading paragraph (Lead), and Text. We chose the TEI format as a target format for the collection namely because of the following three reasons (1) The format is standardized and widely recognized by the community of corpus linguistics, (2) We use it in the ParCzech (Kopp et al., 2021) and ParlaMint (Erjavec et al., 2022) projects compiling parliamentary data into corpora, and thus we can directly use the existing scripts for e.g., linguistic processing, and (3) The data can be visualized and queried in the TEITOK web service.

Most of the item values were converted into TEI XML elements' values without any subsequent modification. The Lead and Text items contain not only the story itself but an HTML code of the original page including Javascript codes as well. We normalized sequences of spaces and removed/replaced problematic Unicode characters. Further, we removed text formatting (e.g., bold text) because it would subsequently make linguistic processing difficult, mainly tokenization. All the scripts are available in the GitHub repository `https://github.com/ufal/media-irozhlas`. The linguistic processing scripts use the API of LINDAT services UDPipe[5] and NameTag[6] for morphological and syntactic analysis (incl. tokenization and lemmatization) and named-entity recognition, resp.

For internal purposes, we uploaded the iRozhlas collection into TEITOK which is an online system for (1) making corpora available and searchable, and (2)

[3] `https://www.mvcc.edu/learning-commons/pdf/signal-phrase-guide-library-and-learning-commons.pdf`

[4] `https://www.irozhlas.cz/`

[5] `https://lindat.mff.cuni.cz/services/udpipe/`

[6] `http://lindat.mff.cuni.cz/services/nametag/`

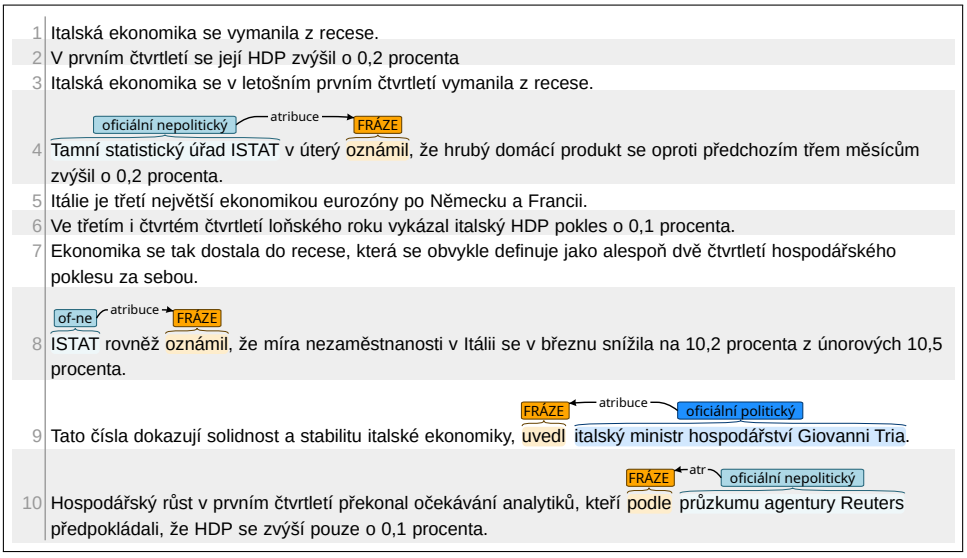

Figure 1: Example of an annotated text in the Brat tool

editing, annotating, and correcting corpora.[7] The files in TEITOK can contain not only the corpus text, but a wide range of annotations, including the annotation by UDPipe and NameTag.

4. Annotation Task

We organized an annotation task to create a gold data set for the task of source detection and classification. In the future, we will address this task automatically using a combination of rule-based approach and machine learning and the gold data set will serve as a train and evaluation data set for any method that we will apply.

section	# of stories
Czech Republic News	272
World News	246
Business	230
Sports	281
Culture	232
Science & Technology	232
Commentary	224
Style	230
Total	1,947

Table 1: Sections in the iRozhlas annotation collection

Annotation collection We specified the following requirements to select the stories from the iRozhlas collection for the annotators:

- Each annotator annotates at least one article from each section (see Table 1).
- The amount of annotation work should correspond to two hours of annotation, which (based on prior tests) corresponds to annotation of a text of about 3,500 words.
- We want to explore the agreement between annotators.

To meet these criteria, we created a subset of the iRozhlas collection to be annotated:

- Each annotator was assigned a unique article from each section with word counts ranging from 200 to 550 words.
- To measure inter-annotator agreement, each annotator was assigned an additional article from the next annotator's folder. If the article selection from the previous steps resulted in less than 2,800 words, we chose the longest one, otherwise the shortest one. In total, 220 files were selected to be annotated by two annotators.

The iRozhlas annotation collection contains 1,947 stories and each annotator was assigned either 9 or 10 stories (9.76 on average).

Tool The Brat[8] editor proved to be the most suitable for our annotation task, especially we appreciate its easy configuration for a large number of annotators and its user friendly GUI for inexperienced annotators. It is a server-client tool with a client-side implemented in

# of annotators	222
# of files to annotate	2,167
# of unique stories in the collection	1,947
# of files with at least one annotation	1,874
# of folders with at least one annotated file	204
# of annotated signals in the files	11,012
# of annotated sources in the files	9,843
# of annotated attribution links in the files	10,110

Table 2: Overall statistics on the annotation

[7] https://lindat.mff.cuni.cz/services/teitok/

[8] https://brat.nlplab.org/

	2017	2018	2019	2020	2021	2017-2021
Czech Republic News	27.0	30.0	24.5	28.5	36.4	29.1
World News	37.4	27.8	33.2	30.0	25.6	31.6
Business	23.4	29.0	26.1	25.7	24.9	26.1
Sport	14.0	13.3	14.5	13.9	11.4	13.5
Culture	15.3	15.8	18.6	15.3	16.9	16.3
Science & Technology	19.3	18.3	24.6	19.5	21.6	20.4
Commentary	13.9	6.7	9.0	8.0	8.2	9.4
Style	18.7	20.5	21.2	24.0	22.9	21.4
All Sections	20.7	19.8	21.4	20.7	21.1	20.7

Table 3: Number of annotated signals per 100 sentences for the sections in a period of five years

a web browser, so that annotators do not have to install the tool, they only open a link in a web browser of their choice. See Figure 1 for an example of a text annotated in Brat.

Each annotator had their own login name and folder with his/her files (stories) to be annotated: Brat displays a selected story and an annotator performs the annotation steps using mouse moves, hot keys or selection context menu.

Instructions

- Read a story sentence by sentence from beginning to end.
- Whenever you recognize an attribution in a sentence:
  1. mark its signal,
  2. mark its source: mark the longest possible noun phrase with modifiers,
  3. classify the source,
  4. create an attribution link going from the source to the signal. If the source is not mentioned in the sentence, create a link to the last mention of the source in the text preceding the current sentence.

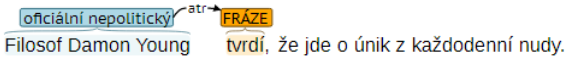

Figure 2: A sample attribution annotation in Brat

Figure 2 illustrates an attribution annotation of the sentence *Philosopher Damon Young claims that this is an escape from the boredom of daily life.*

source class	# of annotations
official-non-political	5,350
official-political	2,404
unofficial	1,215
anonymous-partial	630
anonymous	244

Table 4: Annotations of the source classes

Organization We organized our annotation task as a crowdsourcing annotation. The annotators, bachelor students of the course "Digital Communication and Working with Information",[9] undertook training during a 90-minute lecture. Then each of the 222 annotators received an e-mail with a link to his/her stories in Brat, a login name and a password and a website link to the detailed instructions.[10] This site was being updated with answers to questions raised by the students during the 5 weeks long annotation period. We also discussed these questions with students via e-mail.

5. Annotation Evaluation

During the annotation period, we regularly checked the number of annotated signals and sources and we encouraged students with no annotation to start their task. After its end, there were 286 (out of 1,947) stories with no annotation. This does not necessarily mean that the students did not read them. Undoubtedly there may be stories with no attribution in the collection. The total number of files (incl. the double-annotated stories) used for evaluation was 2,167. As can be seen in Table 2, 11,012 signals and 9,843 sources were annotated in total.

[9] https://is.cuni.cz/studium/predmety/index.php?do=predmet&kod=JKB003

[10] https://ufal.mff.cuni.cz/anotace-citacnich-frazi-v-datech-irozhlas

section	# of attribution links per 100 sentences
Czech Republic News	30.9
World News	34.3
Business	28.9
Sports	14.4
Culture	18.0
Science & Technology	22.5
Commentary	10.5
Style	24.2
Total	22.6

Table 5: Number of annotated attribution links per 100 sentences

	a	a-p	o-n-p	o-p	un
anonymous	1	3	7	6	3
anonymous-partial	–	12	27	19	6
official-non-political	–	–	139	110	11
official-political	–	–	–	177	4
unofficial	–	–	–	–	2

Table 6: Co-occurrence of source classes in the Czech Republic News section

Czech is a typical pro-drop language, which omits the subject if it can be easily reconstructed from the previous context. Therefore we see a difference between the number of the annotated signals and sources. In the following three sentences, we detect one signal *Lukáš Krpálek* and two signals *vysvětluje* (*explains*) and *popisuje* (*describes*), i.e. two attributions: *Mistrem světa se stal Lukáš Krpálek již podruhé. Moje judo je založeno na kondici . . . , vysvětluje. Člověk je musí unavit . . . , popisuje.* Also, a lack of annotators' attention causes some differences.

For each source class, Table 4 displays the number of annotations and Figure 3 shows these numbers for the individual sections. For example, there is a clear evidence that official political sources occur rarely in the Sports and Culture sections.

The average number of annotated attribution links per 100 sentences is 22.6. The Sport section has the lowest attribution "density" (14.4) while the World News section has the highest one (34.3), see Table 5.

The paper (Duffy and Williams, 2011) presents a six-decade longitudinal quantitative analysis on how unnamed sourcing in the Washington Post and The New York Times has changed over time. As for our data, Table 3 displays the number of annotated signals per 100 sentences for each section in a period of five years. For example, we observe with surprise that the number of annotated signals in the CR News section significantly increased while the number of annotated signals in the World News section significantly decreased. A discussion with the news server editor helps to interpret all these data.

Table 6 visualizes co-occurrences of source classes annotated in the Czech Republic News section. Each cell represents two types of sources that appeared in the same article.

annotation	measure	agreement
signals	F1	0.67
sources	F1	0.60
source classes	%	74
source classes	κ	0.58

Table 7: Inter-annotator agreement in recognition of signals, sources and source classes by two annotators

source class	headline (%)	lead+text (%)
official-non-political	36.3	55.0
official-political	26.4	24.3
unofficial	19.7	12.1
anonymous-partial	14.4	6.1
anonymous	3.2	2.4

Table 8: Frequency of source classes in the headlines

Inter-Annotator Agreement was measured on 170 files (not all of the 220 files selected for double annotation contained any annotation in the end), each of which was independently double-annotated by two annotators. Table 7 shows F1 measure for recognition of citation sources and citation phrases, and a percentage agreement and Cohen's kappa for classification of sources recognized by both annotators.

We expected a higher agreement at the beginning of the annotation period. The students are not experienced with this type of tasks and this fact certainly contributed to the given results. But at the same time, we are worried that there is a lack of understanding of what attribution is and how to recognize it in text.

Headlines vs. Leads and Texts Terentieva et al. (2020) studied the attribution technique across headlines in the electronic editions of five leading Spanish mass media outlets between 2010 and 2018. Besides other findings, they concluded that headlines with attribution comprise approximately 15 % of the total number of headlines in the given media. It is perfectly in line with our findings: in our annotated dataset, 13 % of the total number of headlines contain annotation. Table 8 shows how often attribution occurs in headlines and leading paragraphs and texts. It is also interesting to see in Table 9) which signals are the most common: preposition *podle* (*according to*) clearly dominates the leading paragraphs and texts, while it is less frequent in headlines. At the top of the lists there are typically words with neutral polarity. However, in the headlines, there is the verb *varovat* (*to warn*) with negative polarity in order to attract readers.

Authors vs. Sources The analysis of sources for individual authors is interesting. For illustration, we extracted the annotated sources in the articles written by one author. There are 950 such articles (out of 1,947) written by 294 authors. The horizontal axis of the histogram in Figure 4 displays the number of authors and

headline		lead+text	
%	signal	%	signal
19.5	*říkat [to say]*	15.5	*podle [according to]*
7.4	*tvrdit [to claim]*	8.7	*uvést [to state]*
4.3	*říci [to say]*	7.8	*říci [to say]*
3.5	:	4.4	*říkat [to say]*
3.1	*varovat [to warn]*	2.8	*dodat [to add]*
3.1	*podle [accord. to]*	2.4	*informovat [to inform]*

Table 9: Top-6 signals in the headlines and lead+texts

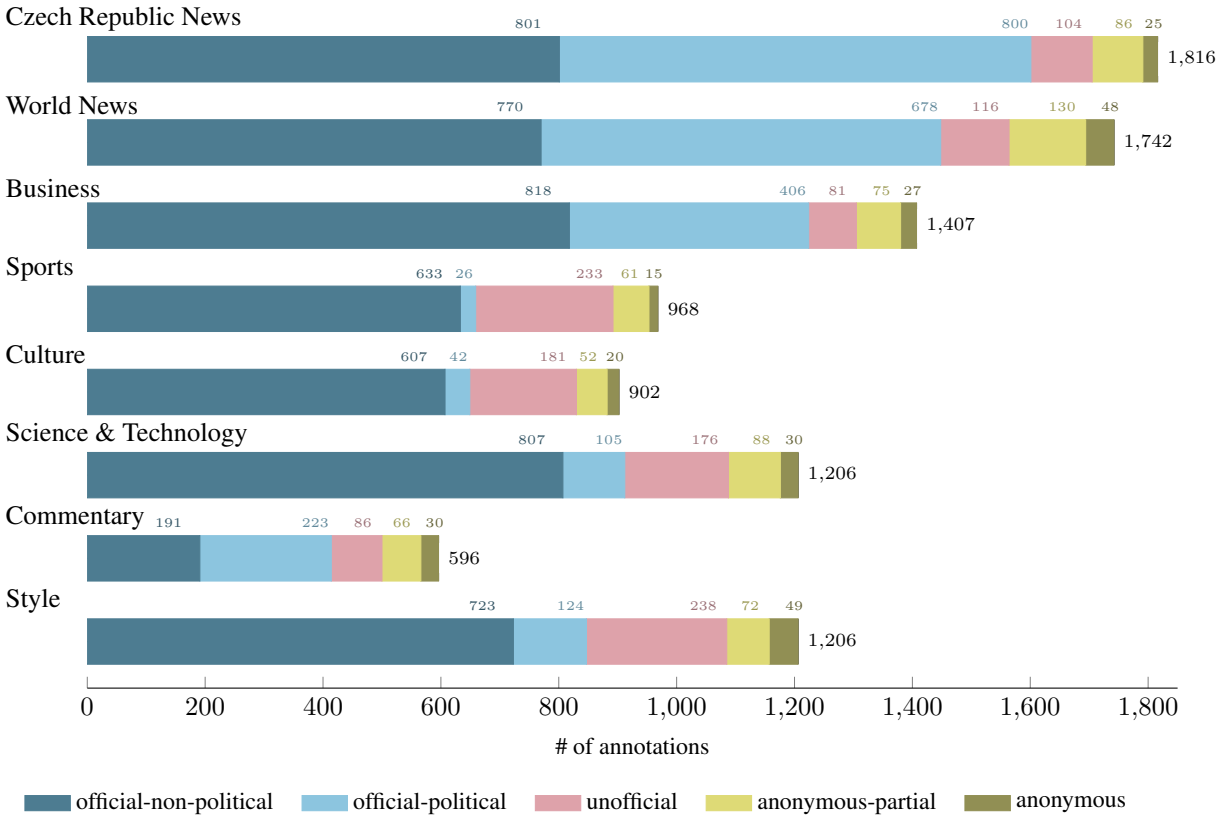

Figure 3: Source classes distribution in the annotated files

the vertical axis represents the average number of annotated sources.

6. Conclusion

Media bias includes, beside others, bias that concerns an analysis of sources attributed in news articles. We focus on sources in the Czech articles published on the iRozhlas web news server being operated by Czech public radio. Namely, we explore the iRozhlas collection of more than 60 thousand articles. In the future, we will perform source detection and classification automatically as a combination of rule-based approach and machine learning. Given that, a very first task is to create a golden data set. Therefore we organized an annotation task. We designed it as a crowdsourcing task that engages typically a large number of individuals to achieve a given goal. In our case, more than 200 bachelor students of the Faculty of Social Sciences, Charles University participated in the annotation. The annotation was one of the course completion conditions and the students were not paid for their work. Based on the annotation analysis presented in this paper we summarize several facts and observations:

First, we set the annotation time to two hours. Then we estimated the length of text to be annotated in two hours and finally we set the number of articles to be annotated by an annotator to 9 or 10. Further, we decided each annotator to annotate articles from all the newspaper sections (see Table 1). Since the students had very little experience with text annotation we did not apply any other criteria for file selection. In total, 1,947 stories from the iRozhlas server were randomly assigned to the students; including the double-assigned files for measuring the inter-annotator agreement, the resulting collection contains 2,167 files.

We set the annotation period to continuous five weeks. An e-mail helpline was active during this period to discuss any topic related to the annotation. Only 5 % of the students took advantage of this opportunity and they typically asked questions of a technical nature.

Once the annotation period ended, we extracted the annotated signals and sources, lemmatized them and represented them as a frequency list.[11] To check how the annotators understand the task, we checked the low-frequency items in this list. No doubt some mistakes are due to the annotators' inattention, but the others show that some students do not recognize attributions in texts at all. This leads us to organize this annotation task again next year and split the annotation period into two parts. The annotation evaluation after the first part will show annotator agreement that we can use in the process of file assignment.

We will focus on more rigorous evaluation of the annotation task using statistical hypothesis testing. We will discuss its results with journalists and news editors.

[11] https://ufal.mff.cuni.cz/anotace-citacnich-frazi-v-datech-irozhlas

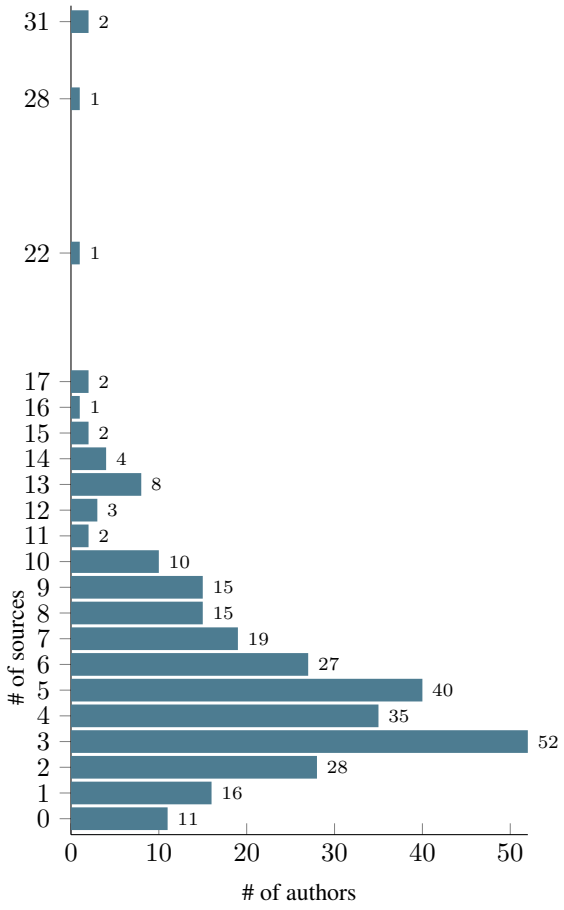

Figure 4: Average number of sources annotated in the articles written by one author

Based on the annotated signals and sources, we will generate queries in e.g., the Corpus WorkBench Query Language that enables searching data analysed by UDPipe and NameTag in TEITOK. For illustration, the query `[form="podle"] <name_type="PER"> []* [xpos="....2.*"][]*</name_type> within s` searches for the signal *podle* (*according to*) and sources being persons in the genitive case.

Acknowledgement

We would like to thank the students for their efforts and Eva Hajičová for her valuable comments on this article. This work was supported by the Technological Agency of the Czech Republic (grant number TL05000057). This work has been using language resources and tools developed and/or stored and/or distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101).

7. Bibliographical References

Baker, B. (1996). *How To Identify, Expose And Correct Liberal Media Bias*. Media Research Center, Alexandria, VA.

Becket, C. (2019). New powers, new responsibilities. `https://drive.google.com/file/d/1utmAMCmd4rfJHrUfLLfSJ-clpFTjyef1/view`. Online.

Duffy, M. J. and Williams, A. E. (2011). Use of unnamed sources drops from peak in 1960s and 1970s. *Newspaper Research Journal*, 32(4):6–21.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., de Macedo, L. D., Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevičius, V., Krilavičius, T., Darǵis, R., Ring, O., van Heusden, R., Marx, M., and Fišer, D. (2022). The ParlaMint Corpora of Parliamentary Proceedings. *Language Resources and Evaluation*. `https://link.springer.com/article/10.1007/s10579-021-09574-0`.

Hamborg, F., Donnay, K., and Gipp, B. (2019). Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

Kopp, M., Stankov, V., Krůza, J. O., Straňák, P., and Bojar, O. (2021). ParCzech 3.0: A Large Czech Speech Corpus with Rich Metadata. In *24th International Conference on Text, Speech and Dialogue*, pages 293–304, Cham, Switzerland. Springer.

Lazarski, E., Al-Khassaweneh, M., and Howard, C. (2021). Using nlp for fact checking: A survey. *Designs*, 5(3).

Leppänen, L., Munezero, M., Granroth-Wilding, M., and Toivonen, H. T. (2017). Data-driven news generation for automated journalism. In *INLG*.

Marconi, F. (2020). *Newsmakers: Artificial Intelligence and the Future of Journalism*. Columbia University Press.

Park, S., Kang, S., Chung, S., and Song, J. (2009). NewsCube: Delivering Multiple Aspects of News to Mitigate Media Bias. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 443–452, New York, NY, USA. Association for Computing Machinery.

Prasad, R., Dinesh, N., Lee, A., Joshi, A., and Webber, B. (2006). Annotating attribution in the Penn Discourse TreeBank. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 31–38, Sydney, Australia, July. Association for Computational Linguistics.

Terentieva, E., Khimich, G., and Veselova, I. (2020). The analysis of citation in headlines in the spanish press. *Heliyon*, 6(1):e03155.

Thorne, J. and Vlachos, A. (2018). Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.