



Title	Citation Estimation Method Using Abstracts of Research Data Articles: A Focus on Scientific Data
Author(s)	Kai, Naoto; Yoshihisa, Tomoki; Shimbaru, Toshiki et al.
Citation	Advances on P2P, Parallel, Grid, Cloud and Internet Computing (Lecture Notes on Data Engineering and Communications Technologies). 2023, 189, p. 1-13
Version Type	AM
URL	https://hdl.handle.net/11094/93143
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

<https://ir.library.osaka-u.ac.jp/>

The University of Osaka

Citation Estimation Method Using Abstracts of Research Data Articles: A focus on *Scientific Data*

NAOTO KAI^{†1} TOMOKI YOSHIHISA^{†2} TOSHIKI SHIMBARU^{†3}
HIDETO YANO^{†1} HIDEYUKI TANUSHI^{†1}

¹ Osaka University, Osaka, Japan

² Shiga University, Shiga, Japan

³ Seinan Gakuin University, Fukuoka, Japan

Abstract: With the trend of open science, efforts have been made to openly utilize research data. Considering the use of shared research data for interdisciplinary research, developing a researcher-friendly abstract writing method in different research fields is pertinent. In this study, we focus on abstracts from *Scientific Data*, a journal specializing in research data. We examine the influence of each part of speech on the utilization of research data through multiple regression analysis of the number of occurrences of the part of speech, the number of words and index-keywords in the abstract, and the number of citations research data article. Based on these results, we set the explanatory variables as the number of nouns, verbs, the other parts of speech, words, and index-keywords in the abstract. Thereafter, we developed a classifier to estimate the number of citations using machine learning. An analysis of the relationship between the number of citations and index keywords was also conducted. We aim to prompt discussions on relevant considerations when writing abstracts.

Keywords : Research data article, Abstract, Number of citations, Scientific Data

1 Introduction

The amount of research data is increasing daily, and the global trend toward open science has made research data management, including storage and distribution, a critical concern. In Japan, the importance of research data management has been recognized, and various efforts have been made to address it. In research data management, there is a need for an ecosystem cycle that extends not only to the release of research data but also to the use of the released research data and the evaluation of the researchers who generated the research data.

While the utilization of research data is expected to promote research efficiency and the fusion of different fields, it is not easy for researchers who do not specialize in a particular research field to understand the outline and value of research data. This is because research data are closely related to specific research fields. Therefore, to promote the use of research data, metadata (particularly data descriptions) that

facilitate the understanding of research overviews are important. Data description, in particular, is important metadata for explaining various research data; however, there has not been sufficient discussion on what kind of description method is effective for the utilization of research data.

Under these circumstances, data repositories and data journals have proliferated worldwide in recent years in response to requests from funding agencies for disseminating research data. Some of the most representative examples are “Edinburgh DataShare,” one of the world’s first data repositories, “Scientific Data” and “Data in Brief,” two data journals that are not limited to a specific field, “Earth System Science Data,” “Earth System Science Data,” and “Chemical Data Collections”. [1] [2] [3] [4] [5] [6] While data journals require researchers to write abstracts according to specific rules, such as word count and index-keywords, as in regular journals, in data repositories, the method of writing abstracts is often left to the researchers. As mentioned above, the abstract of a research data article is an important element that promotes the understanding of the research outline and influences the number of citations of the research data article. We have focused on this in our previous studies. [7] [8] [9]

2 Purpose of This Study

Given the current situation, we assume that it is possible to predict the increase in the number of citations of research data articles through abstracts and attempt to estimate the number of citations using machine learning. In this study, we specifically focus on abstracts from *Scientific Data* journal. First, we consider the factors that affect the number of citations of research data articles. We conduct multiple regression analysis to develop a classifier to estimate the number of citations of research data articles based on the number of occurrences of English parts of speech, the number of words, and the number of index-keywords in the abstracts. Based on these results, we use machine learning to predict the number of citations by setting the explanatory variables as nouns, verbs, other parts of speech, the number of words, and the number of index-keywords and setting the objective variable as the number of citations.

We hope that the results of this study will lead to a method of writing abstracts that will help increase the number of citations, that is, to extract the information that researchers are looking for.

3 Related Research

3.1 Research on the importance of abstracts

Jinfang Niu [10] stated that insufficient data description (a type of metadata equivalent to an abstract in a research data journal) hinders the promotion of data reuse. The reason for the lack of data descriptions is the differing interests of data creators and those who want to use the data which dampen the data creators’ willingness to provide information. Additionally, reduced communication makes it

necessary to prioritize necessary information. Therefore, the tacit information of data creators is given relatively low priority, which hinders the promotion of reuse. In this context, Jinfang's suggestion of the relationship between tacit information and the number of citations of research data is related to this study, in which the number of citations is inferred from information such as the parts of speech that make up the abstract.

3.2 Researching the impact on the thesis abstract

Wong et al. [11] demonstrated that the combined summarization performance was better than the single summarization performance. The surface features they focused on included the position of the sentence in the document and the number of words in the sentence. The content features they focused on were the frequency of occurrence. Resultantly, the summary performance focusing on the surface features had an accuracy of approximately 48%, whereas that focusing on the content features had an accuracy of approximately 40%, indicating that the summary performance focusing on the surface features was better. The combined summary performance of the two features had a precision of approximately 57%, indicating that the combination of the two features improved the summary performance. As this study analyzed the impact of abstracts on the number of citations, it is similar to the study by Wong et al. [11]. However, Wong et al. [11] did not focus on the parts of speech, and their analysis method was different from that used in this study.

3.3 Research on identifying trends through the analysis of abstracts

Matsumoto et al. [12] used article abstracts as data and clarified trends in nursing research in palliative care through text mining analysis, such as investigating the relationships between journal publication year and the term co-occurrence networks. Koyama et al. [13] conducted a quantitative text analysis of abstracts from annual scientific meetings to investigate the characteristics of rehabilitation practices. Although many studies have been conducted to understand article trends, such as frequently occurring words in abstracts, they have not focused on the impact of abstracts on the article readers.

4 Method

4.1 Selection of target data

Scopus, a literature search tool, was used in this study. We included 721 research data articles published in *Scientific Data*, an open-access journal dedicated to the detailed presentations of datasets, that were cited at least once. *Scientific Data* is a research data journal covering all areas of science, launched by the Nature Publishing Group in May 2014. The Group aims to promote the widespread use of research data and give

credit to researchers who publish their data. The journal uses a new format, called a Data Descriptor, as a type of article, which is described in this study as a research data article.

In this study, we specifically included research data articles published from May 2014, when the first issue was published, to 2018. Research data articles, like most academic papers, consist of an abstract at the beginning, a main text describing the research data in detail, and references. In particular, the abstract provides a concise description of the content of the academic paper and is important for judging whether or not to proceed with the main text. The morphological analysis of the abstracts of these 721 research data articles was performed using KHcoder, a free software for quantitative text analysis and text mining developed by Higuchi [14]. A Stanford log part-of-speech tagger [15] was used as the morphological analysis engine. The morphological analysis can show that the frequency of the occurrence of the parts of speech influences the readability of the text. For example, the frequent use of adjectives and adverbs may increase the ambiguity of the text. However, the frequent use of these parts of speech may decrease the restrictiveness of the text and attract the attention of more researchers. Nouns, the most frequently used words in general texts, are important words that are likely to influence the amount of information that a reader receives.

First, the abstracts of the research data articles included in the analysis were selected through 2018 for the following reasons. From Figures 1 to 4, it can be inferred that the number of citations in 2014 and 2015 is the dividing line of whether or not its increase slows after five years since the publication of the research data article. In particular, it can be inferred that the turning point is whether the number of citations exceeds 10 or not. Therefore, we consider 460 research data papers with more than 10 citations as positive examples and 261 research data papers with fewer than 10 citations as negative examples. Table 1 shows the number of positive and negative examples for each year of the research data articles considered in this study.

Table 1 Number of citations for research data articles by publication year

Publication year	Positive example	Negative example	Total
2014	34	14	48
2015	50	23	73
2016	75	43	118
2017	135	59	194
2018	166	122	288
Total	460	261	721

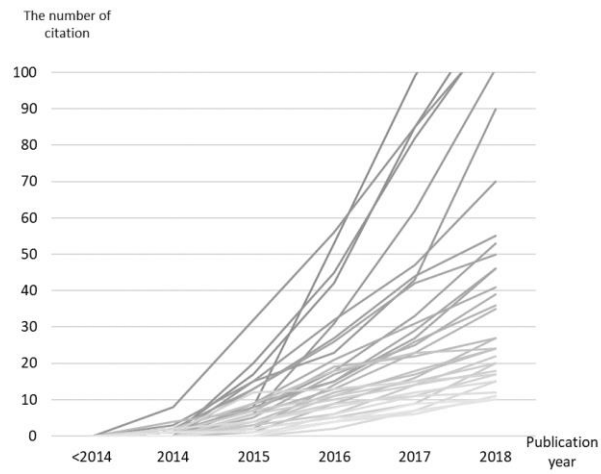


Fig 1. Number of citations (10 or more) for the research data articles published in 2014

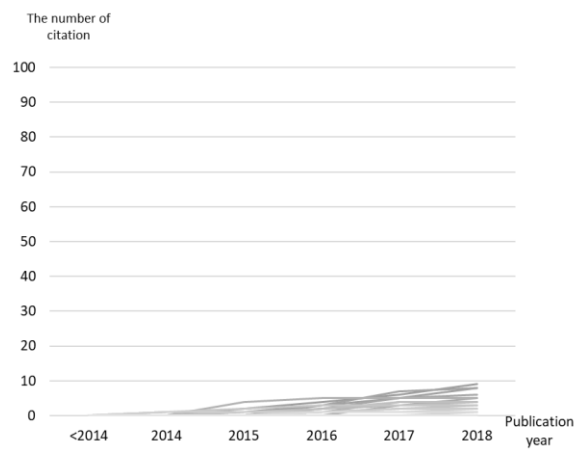


Fig 2. Number of citations (less than 10) for the research data articles published in 2014

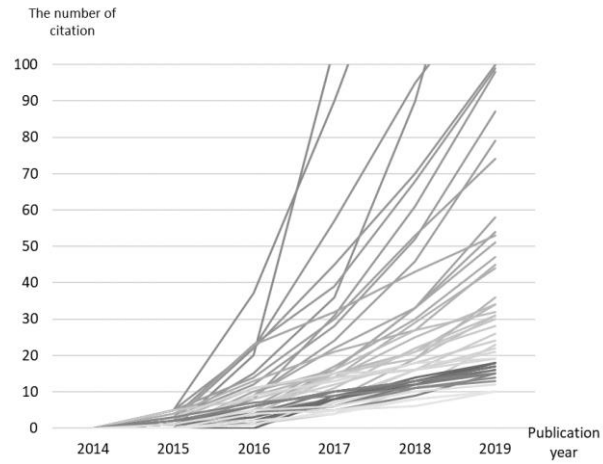


Fig 3. Number of citations (10 or more) for the research data articles published in 2015

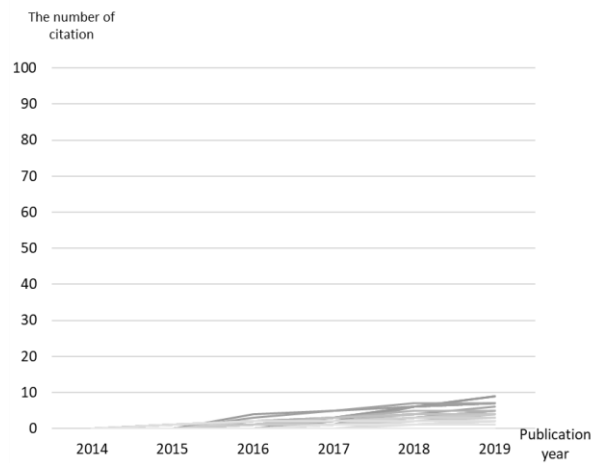


Fig 4. Cumulative citations in 2015 (less than 10)

4.2 Analysis method

First, morphological analysis was used to determine the number of occurrences of nouns, proper nouns, verbs, adjectives, and adverbs in each abstract. Further, conjunctions and prepositions were identified; however, they were excluded from the parts of speech used in this study because they occurred less frequently than nouns, proper nouns, verbs, adjectives, and adverbs.

Multiple regression analysis (5% significance level) was performed using R version 4.2.2 (R Core Team 2022) to estimate the number of citations using machine learning. In addition to the parts of speech classified by the morphological analysis, the number of words in the abstract and the number of index-keywords were added as

explanatory variables to extract factors that significantly impacted the number of citations.[16] Google Colab, which allows Python to be written and executed in a web browser, was used as the execution environment for the machine-learning identification experiments. As listed in Table 2, the training and test data were split in a 7:3 ratio for the 721 datasets.

Table 2. Training and test data

	Training data	Test data
Positive example	317	143
Negative Example	187	74

Additionally, discrimination performance was obtained using machine learning classification models, such as Support Vector Machine (SVM, LinearSVM), K-Nearest Neighbor, Perceptron (Perceptron), Multilayer Perceptron (MLP), naïve Bayes, XGboost (xgb), Logistic Regression (LR), Decision Tree, and Random Forest (Random).

5 Result

5.1 Estimation of factors affecting the number of citations

Table 3 presents the results of the multiple regression analysis (5% significance level). The number of citations increased for the research data papers with a relatively high occurrence of nouns (t-value = 3.182, p-value = 0.002) and words (t-value = 4.418, p-value = 0.000). Further, it increased for research data papers with relatively low occurrences of verbs (t-value = -2.035, p-value = 0.043) and index-keywords (t-value = -2.646, p-value = 0.008).

Table 3. Multiple regression analysis results in R

Explanatory variables	Estimate	Standard error	t-value	p-value
Nouns	0.0057	0.0018	3.1820	0.0015
Proper nouns	0.0035	0.0026	1.3350	0.1822
Adjectives	0.0049	0.0030	1.6600	0.0974
Adverbs	0.0041	0.0069	0.5870	0.5573
Verbs	-0.0064	0.0031	-2.0350	0.0422
Words	0.0026	0.0006	4.4180	0.0000
Index-keywords	-0.0056	0.0021	-2.6460	0.0083

5.2 Discrimination performance by each model of machine learning

The results of the multiple regression analysis described in the previous subsection revealed the factors that significantly impacted the number of citations. Therefore, an experiment was conducted to identify the number of citations, as an objective variable,

by machine learning using nouns, verbs, other parts of speech, and the number of words and index-keywords as explanatory variables.

Table 4 lists the precision, recall, and accuracy values obtained using the best parameters for each model. The F-measure, harmonic mean of goodness-of-fit, and recall were used as evaluation indices. It was found that the SVM (LinearSVM), MLP, xgb, LR, and Random classification models could discriminate the number of citations with F-measure values above 80%. Among these models, the MLP model had the highest F-value of 0.853 and the highest percentage of correct answers (0.788) when estimating the number of citations, indicating that machine discrimination can classify sufficiently high F-values and percentages of correct answers.

Table 4. Estimation results of the number of citations by each model

Models (Parameters)	Parameter values	Precision	Recall	F-measure	Accuracy
SVM (C)	0.1	0.696	1.000	0.821	0.696
LinearSVM (C)	0.05	0.711	0.927	0.805	0.687
KNN (n_neighbors)	3	0.747	0.821	0.782	0.682
Perceptron	—	0.692	0.536	0.604	0.512
MLP (hidden_layer_sizes)	(5, 2)	0.822	0.887	0.853	0.788
NB	—	0.724	0.887	0.797	0.687
xgb (n_estimators)	z	0.738	0.954	0.832	0.733
LR (C)	20	0.714	0.927	0.807	0.691
tree	—	0.695	0.861	0.769	0.641
Random	—	0.716	0.967	0.823	0.710

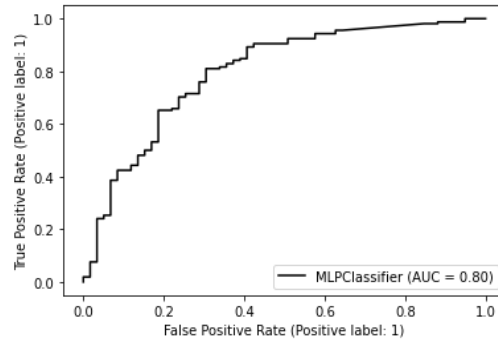


Fig 5. Discrimination performance with MLP

5.3 Relationship between the number of citations and index-keywords

Based on the analysis results from the previous section, it was observed that the higher the frequency of nouns, which are parts of speech that directly convey

information, the higher the number of citations. Conversely, the lower the number of index-keywords, the higher the number of citations. It is commonly observed that index-keywords consist mainly of nouns, and there is a high probability that terms common to both the main text and the abstract are used as index-keywords. Therefore, we investigated whether index-keywords include terms that appear in both the main text and the abstract, and considered the possibility that these index terms influence the number of citations.

First, we examined the presence of index-keywords in both the main text and the abstract. As shown in Tables 5 and 6, we analyzed the top 5 research data papers with the highest number of citations and the bottom 5 research papers with the lowest number of citations. These tables show the proportion of index keywords that are common to both the main text and the abstract among all index-keywords.

Table 5. Proportion of index-keywords common to both main text and abstract (Top 5 Research data papers)

	Published year	Number of citation	Number of index keyword	Index keyword in text and abstract	Ratio of (b) to (a)
1	2017	991	7	5	0.71
2	2017	846	16	2	0.13
3	2020	566	2	1	0.50
4	2018	525	12	2	0.17
5	2014	485	14	1	0.07

Table 6. Proportion of index-keywords common to both main text and abstract (Bottom 5 Research data papers)

	Published year	Number of citation	Number of index keyword	Index keyword in text and abstract	Ratio of (b) to (a)
1	2014	1	18	0	0.00
2	2016	1	27	5	0.19
3	2016	1	14	7	0.50
4	2017	1	2	1	0.50
5	2017	1	15	1	0.07

Table 7. Index-keywords appearing in both main text and abstract (Top 5 Research data papers)

Index keyword					
1	boundary layer	climate	precipitation	prediction	species distribution
2	glioma	Glioma			
3	climate				
4	human	pathology			
5	Molecular				

Table 8. Index-keywords appearing in both main text and abstract (Bottom 5 Research data papers)

	Index keyword						
1	-						
2	graphite	phosphorus	spectroscopy	Graphite	Phosphorus		
3	fresh water	biodiversity	fish	river	Biodiversity	Fresh Water	Rivers
4	election						
5	Washington						

Next, we examined the possibility that index-keyword settings affect the number of citations. Using the index-keywords extracted on Tables 7 and 8, we performed a relevance search on the research data papers published in Scopus *Scientific Data*. The aim was to check whether the research data papers corresponding to the index-keywords used as search terms appeared at the top of the search results. The results are shown in Tables 9 and 10. Table 9 shows the analysis results for the top 5 research data papers, while Table 10 shows the analysis results for the bottom 5 research data papers. However, we did not observe specific characteristics such as high-cited research data papers appearing at the top of the search results and low-cited research data papers appearing at the bottom. In addition, it is possible that the high number of citations is due to the fact that the terms adopted as index terms are used in many other papers, indicating that they are representative and important keywords. Therefore, we also considered the number of research data papers retrieved during indexed searches. As a result, for the top 5 papers with the highest number of citations, there were an average of 240 search results for each index-keyword, while for the bottom 5 papers with the lowest number of citations, there were an average of 41 search results for each index-keyword. Although more detailed analysis is needed in the future, it was found that index-keywords in high cited research papers may include representative terms that are associated with a greater number of research papers.

Table 9. Search results using index-keywords (Top 5 Papers)

	Index-keyword	Total number of search results	Relevance ranking
1	boundary layer	9	3
	climate	386	202
	precipitation	82	24
	prediction	145	12
	species distribution	85	20
2	glioma	4	3
	Glioma	4	3
3	climate	386	29
4	human	1277	254
	pathology	14	0
5	Molecular	257	0

Table 10. Search results using index-keywords (Bottom 5 Papers)

	Index keyword	Total number of search results	Relevance ranking
1	-	0	0
	graphite	1	1
	phosphorus	21	5
2	spectroscopy	35	9
	Graphite	1	1
	Phosphorus	21	5
	fresh water	21	5
	biodiversity	120	17
3	fish	92	3
	river	60	19
	Biodiversity	120	17
	Fresh Water	21	5
	Rivers	60	18
4	election	3	0
5	Washington	5	1

6 Conclusion

In this study, we focused on the abstracts of the research data articles published in *Scientific Data* established to promote the publication and use of research data. Further, we estimated the number of citations of the research data articles using the parts of speech, number of words, and index-keywords in the abstracts.

First, each abstract was morphologically analyzed. Thereafter, a multiple regression analysis was conducted using the number of citations of research data articles as the objective variable. This was applied to identify explanatory variables that significantly affected the number of citations based on the number of parts of speech, number of words, and number of index-keywords. The multiple regression analysis revealed that the explanatory variables affecting the number of citations of research data articles were nouns, verbs, number of words, and number of index-keywords. The number of citations increased as the number of nouns and words increased, whereas the number of citations increased as the number of verbs and index-keywords decreased. From this result, it can be concluded that, similar to general texts, more nouns, which are the parts of speech that directly convey information, contribute to an increase in the number of citations. However, the number of index-keywords showed the opposite trend, suggesting that carefully selected index-keywords contribute the most to increasing the number of citations. Future work is required to analyze the correspondence between the nouns used as explanatory variables and the index-keywords.

Based on the results of the multiple regression analysis, we experimented to estimate the number of citations using machine learning by setting the explanatory variables as nouns, verbs, other parts of speech, number of words, and number of index-keywords. The results showed that many classification models could classify F-measure values above 80%. Among them, the MLP classification model had the highest F-measure (0.853) and accuracy (0.788), indicating that machine discrimination could classify sufficiently high F-measure and accuracy.

An analysis focusing on index-keywords was also conducted. Since it was found that the fewer the index-keywords, the more the number of citations tends to increase,

the relationship between index-keywords and the number of citations was analyzed. When research data articles published in *Scientific Data* were searched in order of relevance using index-keywords that commonly appear in the abstract and main text, no characteristics were found in terms of the expected number of citations and the ranking in which the search results. In addition, when index-keywords were used as search terms, we analyzed how many research data papers were associated with those keywords. As a result, we found that index-keywords of research data papers with a large number of citations may use representative terms assigned to more research data papers. As a future issue, it is necessary to analyze how index-keywords that do not appear in the abstract or the main text affect the number of citations.

References

- [1] Data Share, <https://datashare.ed.ac.uk/>, last accessed 2022.12.1.
- [2] Hrynaskiewicz, I., Yoko, S.: Open Access and Open Data Journals to Facilitate Scientific Data Reuse, *Information Management*, 2014, 57–9, pp. 629–640.
- [3] Scientific Data, <https://www.nature.com/sdata/>, last accessed 2022.12.1.
- [4] Data in Brief, <https://journals.elsevier.com/data-in-brief>, last accessed 2022.12.1.
- [5] Earth System Science Data, <https://www.earth-system-science-data.net>, last accessed 2022.12.1.
- [6] Chemical Data Collections, <https://www.sciencedirect.com/journal/chemical-data-collections>, last accessed 2022.12.1.
- [7] Naoto, K.: A Study of Point of View on Inheritance of Tacit Knowledge to Promote Utilization of Research Data, *Research Report of Information Processing Society of Japan. IOT (2022)* 1–6.
- [8] Kai N., Shimbaru T.: Characteristic Analysis of Data Description in Highly Cited Research Data. *IIAI Letters on Institutional Research Vol. 1. (2022)* 10–18.
- [9] N. Kai, T. Yoshihisa, T. Shimbaru, H. Yano, H. Tanushi, : Citation Estimation Method Using Abstracts of Research Data Articles : Using Abstracts of Scientific Data Articles as An Example, *IPSJ SIG Technical Report IOT-60, (2023)* 1-5.
- [10] Jinfang N.: Overcoming Inadequate Documentation. *Proceedings of the American Society for Information Science and Technology Vol 46. (2010)* 1–14.
- [11] Wong, K.F., Wu, M., Li, W.: Extractive Summarization using Supervised and Semi-Supervised Learning. *International Conference on Computational Linguistics, (2008)* 985–992.
- [12] Matsumoto, K., Sakurai, D., Inoue, R.: Trends in Nursing Research on Palliative Care: Content Analysis of Abstracts using Text Mining. *Bulletin of Tokai University Health Sciences, Vol 23. (2018)* 107–112.
- [13] Koyama T., Michimen K.: Rehabilitation of Central Nervous System Disorders: Characteristics of Medical Care. *Rehabilitation Medicine, Vol 58. (2021)* 317–325.
- [14] Higuchi, K.: Quantitative Text Analysis for Social Research: Toward Inheritance and Development of Content Analysis. 2nd edn. Nakanishiya Shuppan Japan (2020).
- [15] Stanford Log-linear Part-Of-Speech Tagger, <https://nlp.stanford.edu/software/tagger.shtml>, last accessed 2022.12.1.
- [16] R Core Team (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>. Last accessed on 2022.12.1.