Factorial Correspondence Analysis Applied to Citation Contexts

Marc Bertin[1] and Iana Atanassova[2]

[1] Centre Interuniversitaire de Rercherche sur la Science et la Technologie (CIRST),
Université du Québec à Montréal (UQAM), Canada,
bertin.marc@gmail.com
[2] Centre Tesniere, University of Franche-Comte, France,
iana.atanassova@univ-fcomte.fr

Abstract. In this paper, we analyze citation contexts and characterize the different sections of scientific articles in terms of the verbs that appear in citation contexts. We have performed Factorial Correspondence Analysis (CA) using the four sections of the IMRaD (Introduction, Methods, Results and Discussion) structure as categories. Our dataset contains about 80,000 research articles published in the six PLOS journals. The results of this approach show that the sections in the rhetorical structure of research articles have very different characteristics when we take into consideration the occurrences of verbs, and more generally, their lexical content. Our results demonstrate a strong relation between verbs used around citations and the positions in the rhetorical structure.

Keywords: Factorial Correspondence Analysis, Bibliometrics, Citation Analysis, Information Retrieval, Citation Context Analysis, IMRaD

1 Introduction

The study of in-text citations is a very old subject, but it still remains an active area of research. This problem interests many researchers and covers several different disciplines and research areas: Bibliometrics, Information science, Sociology of science, Computer science. Despite numerous studies, there do not yet exist automated approaches for the analysis of in-text citations, mainly because of the difficulties in constructing a complete model of the function of citations. As Cronin [6] points out, the need to have a citation theory and citation context analysis has already been investigated (e.g. [11, 14, 15]). But to this day, we still do not have a model explaining the actions of citations. This task is extremely complex because a lot of factors need to be taken into consideration.

In this paper, we address this problem from the point of view of textual statistics. Recent works have shown that the distribution of in-text citations in scientific articles are strongly correlated to their IMRaD structure (see [5]). In addition, other studies investigate this issue using a lexically-based approach. The study of verbs found in citation contexts is an important step towards a better definition of the meaning of citations acts (see [4]). One of the results presented in this work is that the rhetorical sections in scientific papers do not have the same status according the verb frequency distributions. This means that the position of a text segment in the rhetorical structure governs to a large extent the lexical items (verbs) that are used by the author.

In this paper, we report on a set of experiments to classify sentences containing in-text citations, according to their position in the rhetorical structure. We believe that the study of citation contexts implies observing the use of citations in a certain amount of textual data. In this exploratory study, we will focus on this perspective that seems relevant for the understanding of citation acts.

For this purpose, our study uses a multivariate statistical method, namely Correspondence Factorial Analysis (CFA) (see [9, 2, 3]) to propose an analysis of a dataset of about 8,000 textual contexts of bibliographical references (in-text citations). Correspondence analysis is a technical description of contingency tables and is mainly used in the field of text mining (e.g. [12]).

2 Method

In this study, we investigate the relationship that exists between the rhetorical structure of papers and the text structure and more specifically the lexicon. By analysing occurrence frequencies of different lexical items, the main objective of this method is to achieve an optimal projection of the multidimensional system on a factorial plot. The hypothesis that we want to verify can be formulated as follows: the lexical forms present in the contexts of in-text citations are not randomly distributed. They are strongly dependent on the particular positions of the rhetorical structure.

2.1 Dataset

Our dataset consists of six peer-reviewed academic journals published in Open Access by the Public Library of Science (PLOS): six domain-specific journals (*PLOS Biology, PLOS Computational Biology, PLOS Genetics, PLOS Medicine, PLOS Neglected Tropical Diseases*) and *PLOS ONE*, a general journal that covers all fields of science and social sciences. We have processed the entire dataset of about 80,000 research articles published up to September 2013.

We have identified the section structure in each article by analyzing the section titles. All six journals use similar publication models, where authors are explicitly encouraged to follow the IMRaD (Introduction, Methods, Results, and Discussion) structure. As a result, more than 97% of all research articles contain these four section types, although not always in the same order.

We have identified and extracted all textual segments that contain in-text citations. To do this, the text was segmented into sentences and for each of the four sections we have considered the set of sentences containing in-text citations. As a result, we have obtained a total of 3,314,884 sentences, 31.52% out of which belong to Introduction sections, 19.50% to Methods, 14.66% to Results and 34.33% to Discussion sections.

Next, we propose to use these sets to examine the characteristics of citations in the different sections and the ways citations are used according to their position in the rhetorical structure of articles.

2.2 Protocol

We study the sets of sentences from sections that are identified according to the IMRaD structure and the presence of in-text citations. For this purpose, we use two text analysis tools. The first one is an R Commander plugin (see temis [1]) which provides integrated tools of text mining tasks. Corpora can be imported in raw text. The second one, is a python application called IRaMuTeQ [13] which uses the R libraries. These tools were used in order to produce the outputs for the correspondence analysis and tables.

The set of sentences have been split into words and lemmatized: all different forms of a lexical item are identified and associated with the same lexical item. IRaMuTeQ performs stemming from dictionaries, without disambiguation, also called endogenous lemmatization. After lemmatization, we have filtered all verb forms and ranked them by occurrence frequency for each section. This allowed us to produce a map displaying proximity among variables (Lexical vs Rhetorical Structure).

To perform the analysis, we have created a subset of sentences that consists of about 2,000 randomly extracted sentences for each section of the rhetorical structure and for each journal. This amounts to a total of 48,000 sentences for this analysis. As shown in Table 1, this corpus contains 47,714 unique terms, that have 1,569,201 occurrences.

	Introduction	Methods	Results	Discussion	*Total*
Number of terms	327,506	295,798	337,379	608,518	*1,569,201*
Number of unique terms	21,389	22,848	22,316	28,694	*47,714*
Percent of unique terms	6.5	7.7	6.6	4.7	*3.0*
Number of hapax legomena	8,809	10,539	9,326	11,689	*18,082*
Percent of hapax legomena	2.7	3.6	2.8	1.9	*1.2*
Number of words	327,506	295,798	337,379	608,518	*1,569,201*
Number of long words	124,618	102,422	114,858	222,373	*564,271*
Percent of long words	38.1	34.6	34.0	36.5	*36.0*
Number of very long words	43,499	34,032	38,668	76,932	*193,131*
Percent of very long words	13.3	11.5	11.5	12.6	*12.3*
Average word length	5.7	5.5	5.4	5.6	*5.5*

Table 1: Vocabulary summary by section

3 Results

We have performed Factorial Correspondence Analysis (CA) using the four sections of the IMRaD structure as categories. The interpretation of this analysis is a graph representation of associations between rows and columns. Columns express the sections while the rows correspond to all forms of occurrences. As word meanings strongly depend on their contexts, we consider only sentences containing in-text citations, thus limiting the possible ambiguities.

The interpretation of an axis in CA in a linguistic context is defined by the opposition between the extreme points. Figure 1 presents the projection of the four sections. This figure shows that, for example, the Methods section is in opposition to all other sections. Similarly, the Results and the Introduction sections are opposed to each other on the vertical axis.

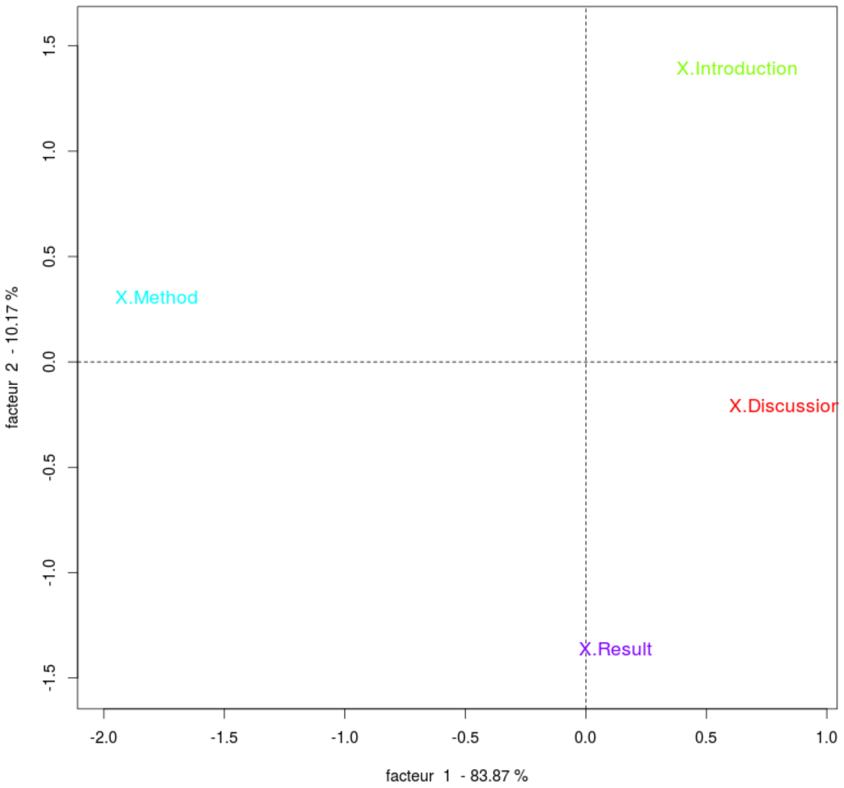

Fig. 1: Correspondence Analysis - Factor 1/2: Projections of Sections on a Factorial Plane

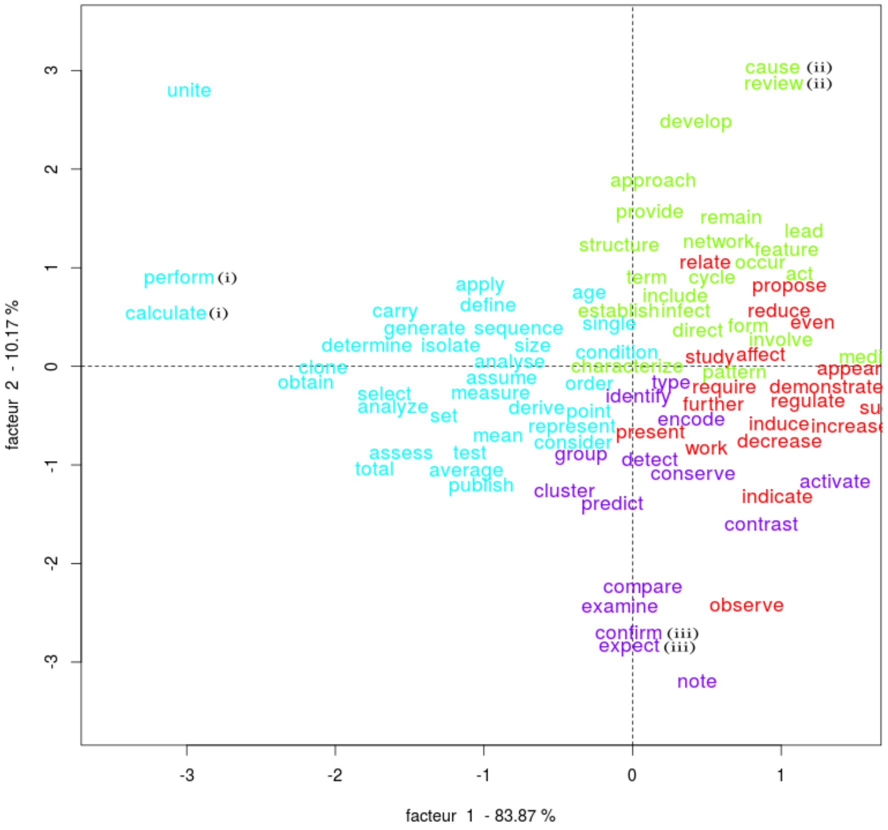

Fig. 2: Correspondence Analysis - Factor 1/2 - without recovery : Projections of Most Frequent Verbs on a Factorial Plane

Figure 2 presents the projection of a sample of the most frequent verbs. For the same verbs, table 2 presents the values for the relative frequencies in the different sections: Introduction - Methods - Results - Discussion. If certain verbs are more or less homogeneous among the sections, some of the verbs, as observed, are only predominantly found in the Discussion and Results sections.

For example, we can see on figure 2 that the verbs **performs** and **calculate** (see i) are mainly present in the Methods section. For the same verbs, table 2 indicates the values [perform - 52.16] and [calculate - 22.47]. The position of the verbs **cause** and **review** (see ii) on figure 2 show that they are characteristic of the Introduction section. This is confirmed by the values given in table 2, respectively [case - 16.79] and [review - 14.06]. Another example are the verbs

Verbs	Discussion	Introduction	Methods	Results
analyse	6.38	6.33	13.18	7.58
approach	9.18	11.47	8.55	5.08
assume	3.7	3.09	9.29	4.61
calculate (i)	1.32	1.18	**22.47**	3.89
cause (ii)	9.35	**16.79**	3.64	5.08
carry	2.98	4.87	14.65	5.67
characterize	4.38	7.96	2.65	6.81
compare	11.22	8.19	11.95	21.07
confirm (iii)	5.74	2.96	5.11	**9.01**
consider	5.87	5.6	7.42	7.07
contrast	9.18	6.19	2.16	9.18
define	4.93	7.19	13.62	6.86
demonstrate	19.3	13.52	2.26	12.27
detect	10.2	7.69	8.21	10.79
determine	6.04	7.46	24.68	11.98
develop	9.18	15.61	7.33	5.88
encode	7.91	12.29	6.05	13.8
establish	5.61	6.51	4.08	5.54
examine	4.04	3.6	4.72	8.76
expect (iii)	5.57	2.96	3.93	**7.96**
identify	18.02	23.16	19.81	26.03
include	25.67	32.86	22.52	24.84
indicate	8.33	5.55	1.57	7.32
involve	14.71	17.93	3.29	14.18
mean	4.72	3.23	11.36	6.73
measure	7.86	7.24	17.89	10.37
note	7.18	2	3.05	7.74
observe	27.03	11.29	7.08	25.1
obtain	4.76	4.32	31.61	10.33
perform (i)	3.87	3.82	**52.16**	7.87
predict	8.8	7.46	9.98	13.16
present	15.6	11.83	9.09	13.08
propose	12.45	10.06	2.75	5.84
provide	11.01	11.92	9.83	5.88
regulate	12.37	12.42	1.08	11.55
remain	5.65	9.01	3.1	4.99
represent	6.8	5.64	7.96	7.32
reveal	6.89	7.15	1.38	6.52
review (ii)	7.95	**14.06**	2.61	4.27
study	91.72	70.54	46.26	60.69
suggest	40.16	25.67	3.34	19.51

Table 2: Relative Frequency of Verbs

confirm and **expect** (see iii) that mainly belong to the Results section. Their values for this section in table 2 are [expect - 7.96] and [confirm - 9.01].

The results of this approach show that the sections in the rhetorical structure of research articles have very different characteristics when we take into consideration the occurrences of verbs, and more generally, their lexical content. Our results demonstrate a strong relation between verbs used around citations and the positions in the rhetorical structure. In addition, figure 1 shows some proximity between the Results and the Discussion sections, as well as between the Discussion and the Introduction sections.

4 Conclusion

This study confirms the results of previous work (see [4]) around the lexical analysis of citation contexts. It also shows that citation contexts are strongly dependent on the rhetorical structure and this is an important factor for the analysis of citation contexts.

The results can also be considered in the perspective of other studies on the distribution of references [5], according to which the distribution of in-text citations is strongly related to the rhetorical structure of articles. They show, for example, the great specificity of the Methods section, because it has a relatively low frequency of in-text citations. In addition, by considering the most frequent verbs in the different sections, our results imply functionality contexts which are specific to the rhetorical structure. Indeed, this study shows that citation contexts in the different sections can be characterized in terms of their lexical content, and more specifically the verbs that appear near in-text citations. Inversely, the function of citations is strongly related to the rhetorical structure and the position of the citation in the article. Taking into consideration the rhetorical structure is therefore necessary for the analysis of citation acts. By studying the different verbs that are present in citation contexts and their relation to the rhetorical structure, we will be able to determine the semantic relations that authors use when they cite other work.

The results of our study have numerous applications, especially in Information Retrieval ([8, 10]) and Bibliometrics (e.g. [7]). The next step is to improve Information Extraction and the analysis of citation networks. Taking into account these results and analyzing more closely the characteristics of citation contexts is an essential step in the understanding the functions of citations and citation acts.

5 Acknowledgments

We thank Benoit Macaluso of the Observatoire des Sciences et des Technologies (OST), Montreal, Canada, for harvesting and providing the PLOS data set.

References

1. Bastin, G., Bouchet-Valat, M.: RcmdrPlugin. temis, a Graphical Integrated Text Mining Solution in R. The R Journal 5(1), 188–196 (2013)
2. Benzécri, J.P.: L'analyse des données: L'analyse des correspondances. Dunod (1973)
3. Benzécri, J.P.: Correspondence Analysis Handbook. (translated from: Pratique de l'analyse des données, 1. Exposé élémentaire. Dunod, Paris). (1992)
4. Bertin, M., Atanassova, I.: A study of lexical distribution in citation contexts through the IMRaD standard. In: Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval co-located with 36th European Conference on Information Retrieval (ECIR 2014). pp. 5–12. Amsterdam, The Netherlands (April 13 2014)
5. Bertin, M., Atanassova, I., Larivière, V., Gingras, Y.: The distribution of references in scientific papers: an analysis of the imrad structure. In: 14th International Society of Scientometrics and Informatics Conference. International Society for Scientometrics and Infometrics, Vienna, Austria (15-19th July 2013)
6. Cronin, B.: The need for a theory of citing. Journal of Documentation 37(1), 16–24 (1981), http://dx.doi.org/10.1108/eb026703
7. Cronin, B.: Bibliometrics and beyond: some thoughts on web-based citation analysis. Journal of Information science 27(1), 1–7 (2001)
8. Glänzel, W.: Bibliometrics-aided retrieval: where information retrieval meets scientometrics. Scientometrics 102, 2215–2222 (2015)
9. Hirschfeld, H.O.: A connection between correlation and contingency. In: Mathematical Proceedings of the Cambridge Philosophical Society. vol. 31, pp. 520–524. Cambridge Univ Press (1935)
10. Mayr, P., Scharnhorst, A.: Scientometrics and information retrieval: weak-links revitalized. Scientometrics 102(3), 2193–2199 (2015)
11. Moravcsik, M.J., Murugesan, P.: Some results on the function and quality of citations. Social studies of science 5(1), 86–92 (1975), http://sss.sagepub.com/content/5/1/86.full.pdf
12. Morin, A.: Intensive use of factorial correspondence analysis for text mining: application with statistical education publications. Statistics Educational Research Journal (SERJ) pp. 1–6 (2006)
13. Ratinaud, P.: IRaMuTeQ:Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires (2009), http://www.iramuteq.org
14. Small, H.: Citation Context Analysis. Progress in Communication Sciences 3, 287–310 (1982)
15. White, H.D.: Citation analysis and discourse analysis revisited. Applied linguistics 25(1), 89–116 (2004)