# A Study of Lexical Distribution in Citation Contexts through the IMRaD Standard

Marc Bertin and Iana Atanassova

CIRST/UQAM, Quebec, Canada,
bertin.marc@courrier.uqam.ca, iana.atanassova@nlp-labs.org

**Abstract.** In this paper we present a large-scale approach for the extraction of verbs in reference contexts. We analyze citation contexts in relation with the IMRaD structure of scientific articles and use rank correlation analysis to characterize the distances between the section types. The results show strong differences in the verb frequencies around citations between the sections in the IMRaD structure. This study is a "one-more-step" towards the lexical and semantic analysis of citation contexts.

**Keywords:** Content Citation Analysis, Citation Contexts, Bibliographic References, IMRaD, Citation Acts, Lexical Distribution

## 1 Introduction

Citation analysis has been the subject of numerous studies during the last decades and there has been a constant interest in producing a theory of citations. The works of Cronin [5–7], Small [21] and Leydesdorff [11] are among the most important in this domain and showed the importance of this research. Liu [12] and MacRoberts [13] explain some of the difficulty of the task and Mutschke et al. [15] propose a new model for Information Retrieval in scholarly information systems. Teufel et al. [25, 24] propose an annotation scheme for citation functions and discourse-level argumentation. Bertin et al. [4] show a first study of the correlation between the distribution of citations in articles and section types in the IMRaD structure. Their large-scale study examines only the number of citations according to the positions in the text and in the sections but does not rely on further linguistic analyses of the citation contexts.

Scientific articles typically follow the standardized IMRaD (Introduction, Method, Result and Discussion) structure. It gives a rhetorical outline for scientific writing that began to predominate in 1965 and, as Sollaci [23] explains, it was introduced as standard in 1979. During the last decade, many guidelines, surveys and editorial requirements impose this structure throughout scientific literature, especially in the biomedical domain [10, 14, 8]. On the other hand several studies deal with the effects of the use of the IMRaD style [16]. In this paper, we analyze citation contexts in the light of the IMRaD structure, by examining the correlations between the verbs that appear in citation contexts

and the section types. Our hypothesis is that the verbs that appear close to bibliographic citations in texts most frequently define the relation between the article's author and the cited work. Thus, this study contributes to the understanding of the IMRaD structure and the different roles of citations according to their position in articles.

## 2 Method

We have processed a corpus of scientific articles to produce ordered lists of verbs according to their occurrence frequencies in the different section types. Our method is based on the following steps: (i) the XML documents are parsed, sections are extracted and categorized according to the four section types of the IMRaD structure; (ii) we segment the sections into sentences and extract the sentences containing references; (iii) we use POS-tagging and lemmatization tools to identify verbs in citation contexts and construct the ranked verb lists according to their frequencies in each section type.

### 2.1 Dataset

For this study, we have used a corpus of five scientific journals: *PLoS Biology*, *PLoS Computational Biology*, *PLoS Genetics*, *PLoS Neglected Tropical Diseases* and *PLoS Pathogens*, published by PLoS[1] and available in Open Access in the XML format. The articles are structured using the Journal Article Tag Suite (JATS)[2], where the sections in the text are represented as separate elements. We have processed the entire set of research articles of these journals up to September/October 2012. Table 1 shows the number of articles and citation contexts extracted from each journal.

| Journal | Nb of articles | Nb of sentences | Nb of citations | Citation contexts |
|---|---|---|---|---|
| PloS Bio. | 1,587 | 356,816 | 150,429 | 79,703 |
| PloS Comp. Bio. | 1,976 | 487,045 | 177,742 | 92,437 |
| PloS Gen. | 2,435 | 544,569 | 227,121 | 126,230 |
| PloS Negl. Trop. Dis. | 1,240 | 200,920 | 83,402 | 45,714 |
| PloS Path. | 2,208 | 496,371 | 209,685 | 115,750 |
| *Total* | *9,446* | *2,085,721* | *848,379* | *459,834* |

Table 1: Corpus statistics

[1] http://www.plos.org

[2] This Standard is an application of NISO Z39.96-2012 and JATS is a continuation of the NLM Archiving and Interchange DTD (http://jats.nlm.nih.gov)

### 2.2 Section Categorization and Part-Of-Speech-Tagging

Each section is presented in an XML element containing a title and some text content. Our first task was to categorize the sections according to the four types of the IMRaD structure: *Introduction*, *Method*, *Result* and *Discussion*. To do this, we analysed the section titles and used a set of regular expressions related to each section type in order to account for the possible variations in section titles. For example, the Method section type can be expressed by several different titles such as "Method", "Methods", "Method and Model", etc.

The articles in our corpus often contain other section types such as additional information, acknowledgement, etc. that were not taken into consideration. A small number of articles in the corpus do not follow the IMRaD structure and use domain-specific section titles. They were excluded from the study.

The basic unit in our study is the sentence and the quality of the sentence segmentation is important to reliably determine the citation contexts. In fact, rather than define the contexts in terms of number of words around citations, we prefer to use the sentences boudaries as natural delimiters of citation contexts. Sentences are basic linguistic units of texts that we consider as most suitable to model text progression. Table 2 shows the number of citation contexts extracted for each section type.

| Journal | Introduction | Method | Result | Discussion |
|---|---|---|---|---|
| PloS Bio. | 19,769 | 13,911 | 25,263 | 20,760 |
| PloS Comp. Bio. | 25,721 | 18,964 | 27,907 | 19,845 |
| PloS Gen. | 31,476 | 23,239 | 39,781 | 31,734 |
| PloS Negl. Trop. Dis. | 14,103 | 9,611 | 5,533 | 16,467 |
| PloS Path. | 31,107 | 21,202 | 29,676 | 33,765 |
| *Total* | *122,176* | *86,927* | *128,160* | *122,571* |

Table 2: Citation contexts per section

The extracted citation contexts were processed using TreeTagger, a part-of-speech tagger[3] [18, 19] which performs both part-of-speech-tagging and lemmatization. In the output verb forms are tagged by labels such as *VB, VBD, VBG, VBN, VBP, VBZ* that stand for *base form, past tense, present participle*, etc. This allowed us to extract the around 11,000 verb occurrences from the processed sentences.

## 3 Results

Taking into consideration the set of verbs that appear in all four sections, we have obtained a set of 1807 verbs. Then we produced the ranked list of verbs for each section, ordered by the verb frequencies.

[3] http://nlp.stanford.edu/downloads/tagger.shtml

A classical phenomenon is the fact that most of the verb occurrences in citation contexts belong to only a small set of verbs. Table 3 shows that, for example, in the *Introduction* section, 70 verbs account for 50% of all verb occurrences, and 486 verbs account for 90% of the occurrences.

| Percentage | Number of Verbs in Citation Contexts | | | |
|---|---|---|---|---|
| | Introduction | Method | Result | Discussion |
| 10% | 5 | 1 | 5 | 4 |
| 25% | 21 | 3 | 16 | 17 |
| 50% | 70 | 35 | 58 | 59 |
| 75% | 209 | 139 | 184 | 187 |
| 90% | 486 | 368 | 429 | 461 |

Table 3: Distribution of Verbs in Sections

Table 4 shows the ranked lists of the top 10 most frequent verbs for each section type. It is interesting to observe some of the differences. For example, we can see that the verb *show* does not appear in the *Method* section while it is on the first or second position in all the other sections. This means that the verb *show* is used very often in citation contexts except in the *Method* section where it is quite rare. Similarly, we can observe that the *Method* section contains some specific verbs (*perform, follow, obtain, generate*) that do not appear among the top 10 verbs of any other section.

| Rank | Introd. | Method | Result | Discussion |
|---|---|---|---|---|
| 1 | show | use | use | show |
| 2 | use | perform | show | suggest |
| 3 | include | follow | find | use |
| 4 | suggest | obtain | report | report |
| 5 | identify | generate | observe | find |
| 6 | find | base | suggest | include |
| 7 | require | determine | identify | observe |
| 8 | associate | contain | express | require |
| 9 | involve | calculate | see | associate |
| 10 | lead | carry | include | involve |

Table 4: Top 10 of the Most Frequent verbs in the four section types

Figure 1 gives the heatmaps for some selected verbs along the text progression of each section. The horizontal axis corresponds to the progression of the text in each section, from 0% to 100%. Most of these verbs express citation acts. This representation shows that the densities of some verbs vary considerably, especially in the beginnings and ends of the sections. Certain verbs, such as

*perform, obtain, include, describe*, have rather important variations. This result is compatible with the hypothesis that certain citation functions are more likely to be present at some specific positions in texts. From in Information Retrieval point of view, it can be interesting to take this into account for the definition of new term weights related to text positions.

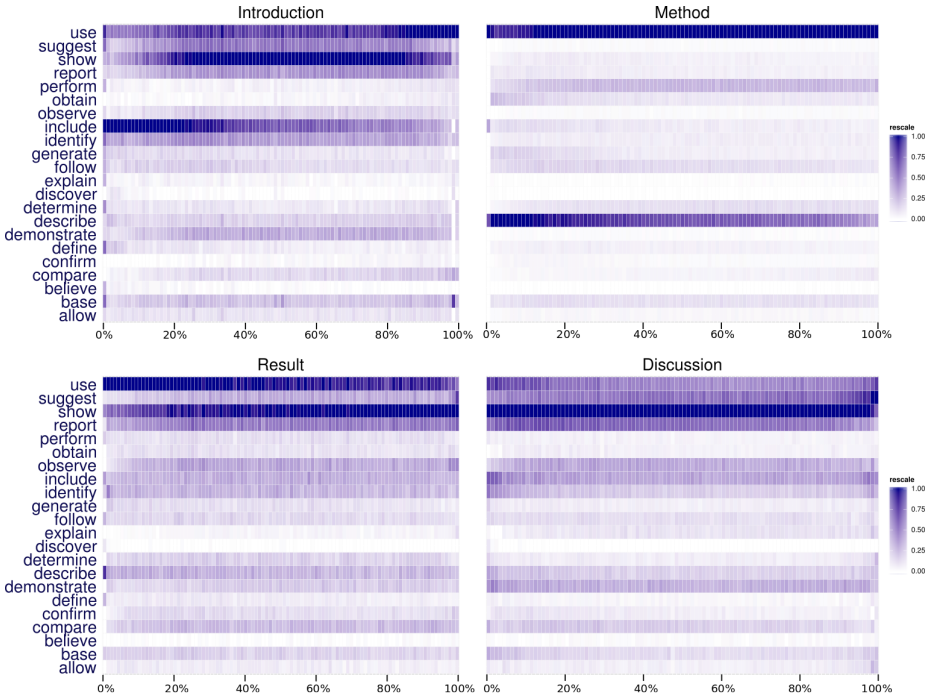


**Fig. 1.** Density of Verbs in Citation Context

To compare and observe the correlations between the different ranked lists, we have used the Kendall tau rank correlation coefficient [9] that provides a measure for the similarity of ordered lists and has an intuitive interpretation.

The Kendall $\tau$ measure is defined as:

$$\tau = \frac{(\mathrm{C}) - (\mathrm{D})}{\frac{1}{2}n/(n-1)}, \tag{1}$$

where $C$ is the number of concordant pairs and $D$ is the number of discordant pairs. $\tau \in [-1, 1]$, $\tau = 1$ if the ranks are identical and $\tau = -1$ if the ranks are inverse.

Figure 2 shows the values of Kendall $\tau$ and the scatterplots for the different section pairs. The scatterplots were obtained by comparing the ranked lists of

verbs for each section pairs. On the horizontal and vertical axes we have the 1807 verbs that appear in all sections. The verbs are arranged according to their rank in the *Introduction* section.3

The biggest similarity is between the *Introduction* and the *Discussion* sections ($\tau = 0.76$), which means that for these two sections the majority of the verbs are ranked on similar positions. On the corresponding scatterplot, this is expressed by the density around the main diagonal. These two sections use most often the same verbs in citation contexts. The similarity between the *Method* and the *Result* is the smallest ($\tau = 0.39$) which means that most of the verbs in the *Result* are rarely employed in the *Method* and vice versa. On the scatterplot this is expressed by a larger dispersion which accounts for the fact that these two sections tend to make use of different sets of verbs around citations. A similar case is the pair *Method* and *Discussion* which also shows large dispersion.

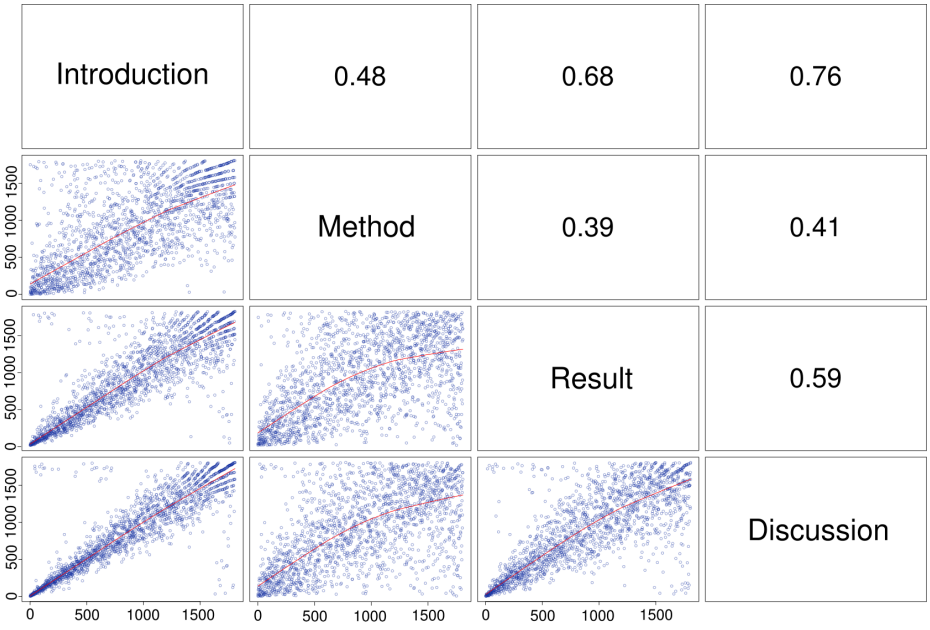


**Fig. 2.** Scatterplots of section pairs and values for Kendall $\tau$

## 4 Conclusion

These results show clearly that the section structure of research papers is an important element to consider as classifiers for citation context analysis. Furthermore, we are able to propose corpora of verb classes related to sections in the IMRaD structure of research papers. These corpora can serve as a reference

data for other works, for example construction of large-scale corpora dedicated to machine learning (see Athar and Teufel [1]), citation-based methods for Information Retrieval (see Ritchie et al. [17]), construction of linguistic resources for semantic annotation (see Bertin [2, 3]), validation of ontologies such as CiTO (see Shotton [20]), validation of frameworks for syntactic and semantic analysis of citation contexts (see Zhang [26]). In a similar perspective, Small [22] proposes to analyse the attitudes and dispositions toward the cited work using cue words in 304 citation contexts. Our study tries to extend this type of approach, by analysing a large number of citation contexts (more than 450,000) and by focusing only on the verbs in order to study the lexical distribution phenomena in relation with the rhetorical structure.

This work confirms the hypothesis that citations play different roles according to their position in the rhetorical structure of scientific articles. The study of citation act verbs is the first step for the categorization of citations and network structures, such as co-citations and bibliographic coupling. Our results show that citation acts are expressed by a relatively small number of verbs that appear in citation contexts. The next step will be automatic semantic reference categorization based on the verbs in the citation contexts as well as other contextual elements.

## References


1. Athar, A., Teufel, S.: Context-enhanced citation sentiment detection. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 597–601. Association for Computational Linguistics (2012)
2. Bertin, M.: Categorizations and Annotations of Citation in Research Evaluation. In: Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference (2008)
3. Bertin, M., Atanassova, I.: Semantic Enrichment of Scientific Publications and Metadata. D-Lib Magazine 18(7/8) (2012)
4. Bertin, M., Atanassova, I., Lariviere, V., Gingras, Y.: The Distribution of References in Scientific Papers: an Analysis of the IMRaD Structure. In: Proceeding of 14th International Society of Scientometrics and Informetrics Conference. International Society for Informetrics and Scientometrics, Vienna, Austria (15th-19th July 2013)
5. Cronin, B.: The Need for a Theory of Citing. Journal of Documentation 37(1), 16–24 (1981)
6. Cronin, B.: The Citation Process. The Role and Significance of Citations in Scientific Communication. London: Taylor Graham, 1984 1 (1984)
7. Cronin, B.: Metatheorizing Citation. Scientometrics 43(1), 45–55 (1998)
8. International Steering Committee of Medical Editors: Uniform Requirements For Manuscripts Submitted To Biomedical Journals. The British Medical Journal 1(6162), pp. 532–535 (1979), http://www.jstor.org/stable/25431277
9. Kendall, M.G.: Rank Correlation Methods. Charles Griffin & Co. Ltd., London (1948)
10. Kucer, S.L.: The making of meaning reading and writing as parallel processes. Written Communication 2(3), 317–336 (1985)

11. Leydesdorff, L.: Theories of Citation? Scientometrics 43(1), 5–25 (1998)
12. Liu, M.: Progress in Documentation the Complexities of Citation Practice: a Review of Citation Studies. Journal of Documentation 49(4), 370–408 (1993)
13. MacRoberts, M.H., MacRoberts, B.R.: Problems of Citation Analysis. Scientometrics 36(3), 435–444 (1996)
14. Meadows, A.: The Scientific Paper as an Archaeological Artefact. Journal of information science 11(1), 27–30 (1985)
15. Mutschke, P., Mayr, P., Schaer, P., Sure, Y.: Science models as value-added services for scholarly information systems. Scientometrics 89(1), 349–364 (2011)
16. Oriokot, L., Buwembo, W., Munabi, I., Kijjambu, S.: The Introduction, Methods, Results and Discussion (IMRAD) Structure: a Survey of Its Use in Different Authoring Partnerships in a Students' Journal. BMC research notes 4(1), 250 (2011)
17. Ritchie, A., Teufel, S., Robertson, S.: Using terms from citations for ir: Some first results. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R. (eds.) Advances in Information Retrieval, Lecture Notes in Computer Science, vol. 4956, pp. 211–221. Springer Berlin Heidelberg (2008)
18. Schmid, H.: Probabilistic Part-of-speech Tagging Using Decision Trees. In: Proceedings of international conference on new methods in language processing. vol. 12, pp. 44–49. Manchester, UK (1994)
19. Schmid, H.: Improvements in Part-of-speech Tagging with an Application to German. In: In Proceedings of the ACL SIGDAT-Workshop (1995)
20. Shotton, D., et al.: Cito, the Citation Typing Ontology. Journal of Biomedical Semantics 1(Suppl 1), S6 (2010)
21. Small, H.: Citation Context Analysis. Progress in communication sciences 3, 287–310 (1982)
22. Small, H.: Interpreting maps of science using citation context sentiments: a preliminary investigation. Scientometrics 87(2), 373–388 (2011)
23. Sollaci, L.B., Pereira, M.G.: The Introduction, Methods, Results, and Discussion (IMRAD) Structure: a Fifty-year Survey. Journal of the Medical Library Association 92(3), 364 (2004)
24. Teufel, S., Siddharthan, A., Tidhar, D.: Automatic classification of citation function. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. pp. 103–110. Association for Computational Linguistics (2006)
25. Teufel, S., Siddharthan, A., Tidhar, D.: An annotation scheme for citation function. In: Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue. pp. 80–87. Association for Computational Linguistics (2009)
26. Zhang, G., Ding, Y., Milojevic, S.: Citation Content Analysis (CCA): A Framework for Syntactic and Semantic Analysis of Citation Content. CoRR abs/1211.6321 (2012)