

Mapping the Linguistic Context of Citations

by Marc Bertin, Iana Atanassova, Vincent Larivière and Yves Gingras

Mapping Science

EDITOR'S SUMMARY

Scientific papers are routinely structured in sections for introduction, methods, research and discussion, a standard since the 1970s. Citations originating within each section serve different purposes and can be meaningfully classified according to position, shedding light on an author's purpose for the citation. Furthermore, words near the citations in the various sections differ, providing the basis for lexical and semantic analysis of citation contexts. Approximately 50,000 scientific papers from seven PLOS journals published between 2009 and 2012 were analyzed for citation use within the identifiable document structure and for verbs used in the context of the citations. Frequencies of verbs in the four section types demonstrate the predominant use of certain words by section. Introduction sections showed greater variety of verbs, while a more limited range of verbs was seen in Methods sections. The lexical distribution process may be applied to other contexts supporting text processing based on XML format.

KEYWORDS

bibliographic citations	verbs
document structure	contextual information
linguistic analysis	intent

Marc Bertin is a postdoctoral fellow at the Centre Interuniversitaire de Recherche sur La Science et La Technologie at the Université du Québec à Montréal, Canada. He can be reached at [bertin.marc<at>gmail.com](mailto:bertin.marc@gmail.com).

Iana Atanassova is an assistant professor at Centre Tesniere at the Université de Franche-Comté, France. She can be reached at [iana.atanassova <at>univ-fcomte.fr](mailto:iana.atanassova@univ-fcomte.fr).

Vincent Larivière holds the Canada Research Chair on the Transformations of Scholarly Communication at the Université de Montréal, Canada. He can be reached at [vincent.lariviere<at>umontreal.ca](mailto:vincent.lariviere@umontreal.ca).

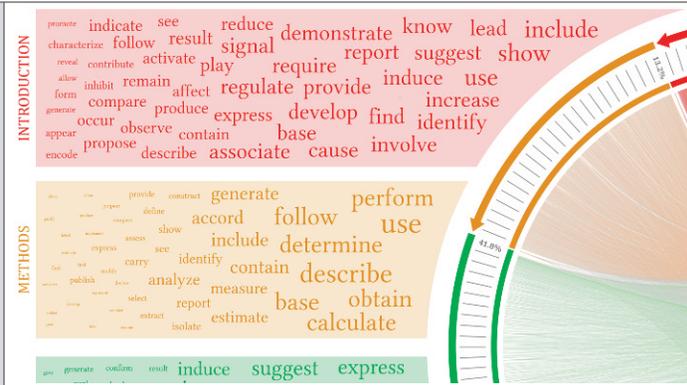
Yves Gingras is a professor in the département d'histoire at the Université du Québec à Montréal, Canada. He can be reached at [gingras.yves<at>uqam.ca](mailto:gingras.yves@uqam.ca).

The IMRaD structure of scientific papers (Introduction, Methods, Results and Discussion) was adopted by most journals since the mid-20th century and became standard in the 1970s. It was introduced to facilitate the reading of publications and to provide faster access to information by standardizing the argumentative structure of articles. The IMRaD sequence provides an outline for scientific writing, dividing the articles into four sections, each one having a specific rhetorical function.

While citations have been studied extensively from a quantitative point of view, our approach is motivated by the need to examine more closely the position of citation references at the level of sentences and observe the trends related to the ways authors cite previous works in different sections of a publication. In this study, we research how authors use citations in the different sections of scientific articles that follow the IMRaD structure. We consider that verbs found in sentences containing citations are an important indicator of the purpose of citations and of the reasons behind citing a given document. Results of the study are communicated using a visualization that shows the relations between verbs used around references and the structure of scientific papers.

Figure 1 presents a cutout of the overall map. The numbers inside the circular representation correspond to the average positions of section boundaries observed in the articles: 13.2% for the end of the Introduction and 41.8% for the end of the Methods section. By showing the differences in the frequencies of verbs that appear in the four section types of the IMRaD structure, this map communicates that the way an author cites a reference depends highly on the position in the IMRaD structure. This study is an important step towards the lexical and semantic analysis of citation contexts. The results indicate that the rhetorical structure of scientific articles determines the positions of references and their relation to the

BERTIN, ATANASSOVA, LARIVIÈRE and GINGRAS, continued

FIGURE 1. Most frequent verbs in the Introduction and Methods sections

article's study. The verbs that we have identified can further be used as clues for the categorization of the relations between authors and for the analysis of citation networks.

Creating a Map of Citation Contexts

The dataset used in this study comprises the entire set of PLOS (Public Library of Science) documents over the period 2009-2012. It contains about 50,000 scientific papers in seven different journals. Most of the articles in the corpus (about 87%) are in the domains of biology and medicine, and the rest of the articles cover a wide range of subject areas such as computer and information sciences, physics and social sciences. The articles are published in Open Access in XML (data harvested from <http://www.plos.org> in October 2012) using the Journal Article Tag Suite (JATS). 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>). Table 1 contains general statistics on the corpus.

Automatic processing of the data proceeded as follows: Firstly, in order to link each citation context with its position in the IMRaD structure, we identified the different sections of the papers (Introduction, Methods, Results and Discussion) and their sizes in terms of number of sentences. In order to do so, we processed the section titles using a set of regular expressions specifically designed for this task. After analyzing the density of citations along the four main section types, we showed that the

distribution of references along the text progression is essentially invariant across the different PLOS journals [1]. This first result showed that a strong relation exists between the IMRaD structure and the use of citations in scientific writing.

Secondly, we extracted verb occurrences found in citation contexts. Verbs were identified using the Stanford POS-tagger [2]. As a result, we obtained the frequency of each verb in the four section types, as well as the links between their occurrences and the positions in the text. These data were visualized using the CIRCOS tool [3], written in Perl, which allows the visualization of information in a circular layout. The data related to the positions of verbs along the text progression were converted into plain text configuration files for the CIRCOS tool. Taking into consideration the verbs that appear in all four sections, we have obtained a set of 1,807 unique verbs. About 500 of them account for 90% of all citation contexts [4]. For the visualization, we have used the top 87 most frequent verbs that account for about 60% of all citation contexts.

Visualization Details and Observations

The graphical representation of the links between verbs in citation contexts and positions in the rhetorical sections can be viewed from two

TABLE 1. General statistics on the PLOS corpus

Journal	Number of articles	Avg number of sentences per article	Avg number of citation contexts per article
PLOS Biology	2,965	141.77	54.63
PLOS Computational Biology	2,107	242.00	87.49
PLOS Genetics	2,560	218.80	91.09
PLOS Medicine	2,228	95.98	39.62
PLOS Neglected Tropical Diseases	1,366	157.44	67.57
PLOS Pathogens	2,354	216.91	93.41
PLOS ONE	33,782	177.90	74.10
All PLOS journals	47,362	178.19	73.55

TABLE 2. Top 10 most frequent verbs in the four section types

Rank	Introduction	Methods	Results	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

different perspectives. On the one hand, it characterizes the four section types through the sets of verbs that are most prominent in each section. For example, we can observe that the Introduction section uses citations in a different way than the Methods section, as the former shows a great diversity in the verbs that are used while in the latter a small number of verbs (*use, describe, obtain, calculate, perform*) stands out as having very high relative frequencies in citation contexts. On the other hand, each verb can be observed independently and analyzed with respect to the sections in which it appears. For example, the verb *show* occurs frequently in all sections except for the Methods section, and the verb *describe* is most frequent in the Methods and Discussion sections and occurs rarely in the other two sections. Table 2 shows the ranked lists of the top 10 most frequent verbs in the four section types. This result confirms that authors use different verbs to introduce citations according to the position in the rhetorical structure and these positions are therefore an important factor for the analysis of citation acts.

Figure 2 represents one of the most frequent verbs with its left context using a textual tree. These contexts were extracted from the corpus that we studied and the tree shows some of the patterns of the use of this verb. Such representations can be used to analyze the different semantic values of the verb according to the contexts in which it appears. For example, the first

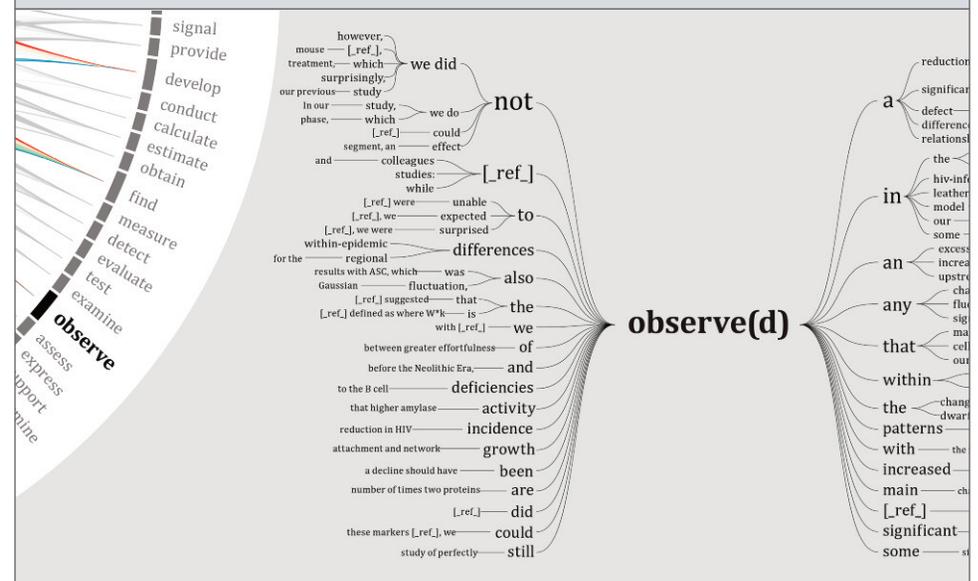
type of context contains “not” and is used by scientists to indicate observations that differ from previous work.

From the point of view of bibliometrics, this map shows clearly that the section structure of research papers is an important element to consider as a classifier for citation context analysis. These results confirm the hypothesis that citations play different roles according to their position in the rhetorical structure of scientific articles. The study of citation through the frequency of verbs used in sentences is a first step towards a better categorization of citations.

Outlook

Our approach is designed around open-source tools such as Stanford POS-tagger and CIRCOS. While we focus on verbs in citation contexts, other linguistic phenomena can be studied using a similar approach, such as sentiment analysis or lexical distributions. However, the availability of the corpus in XML format is an important factor for automatic text processing

FIGURE 2. One of the most frequent verbs and its left context



BERTIN, ATANASSOVA, LARIVIÈRE and GINGRAS, continued

because it gives access to the articles' content in structured full text. Our processing relies on XML-parsing tools that we specifically developed for the JATS schema.

Our current research involves developing textual navigation interfaces

using R-Shiny in order to explore different visualizations of texts. We are also working on the processing of large textual datasets in order to produce the tools for text navigation and analysis that could be used by different communities in Sociology, Semantic Web or natural language processing. ■

Resources Mentioned in the Article

- [1] Bertin, M., Atanassova, I., Larivière, V., & Gingras, Y. (In press). The invariant distribution of references in scientific articles. *Journal of the Association for Information Science and Technology*.
- [2] Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, 12, 44–49.
- [3] Krzywinski, M. I., Schein, J. E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., . . . Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, 19 (9), 1639-45. doi: 10.1101/gr.092759.109
- [4] Bertin, M., & Atanassova, I. (2014). A study of lexical distribution in citation contexts through the IMRaD standard. *Proceedings of the First Workshop on Bibliometric-Enhanced Information Retrieval co-located with the 36th European Conference on Information Retrieval (BIR@ECIR 2014)*, 5-12. (CEUR Workshop Proceedings, 1143). Retrieved from <http://ceur-ws.org/Vol-1143/paper1.pdf>