Factorial Correspondance Analysis Applied to Citation Contexts

Marc Bertin[1] **and Iana Atanassova**[2]

March 29, 2015

[1] CIRST - Université du Québec à Montréal (UQAM),Canada
[2] Centre Tesniere, University of Franche-Comte, France

Bibliometric-enhanced Information Retrieval (BIR 2015)
ECIR 2015, Vienna, Austria

Research Problem

Scientific papers usually follow a specific rhetorical structure: IMRaD (Introduction, Method, Result and Discussion)

- The distribution of in-text citations in strongly correlated to the IMRaD structure.
- The study of verbs in citation contexts is an important step towards a better definition of the meaning of citation acts.

Objectives: Investigate the relationships that exist between the rhetorical structure and the vocabulary used near citation contexts.

Research Problem

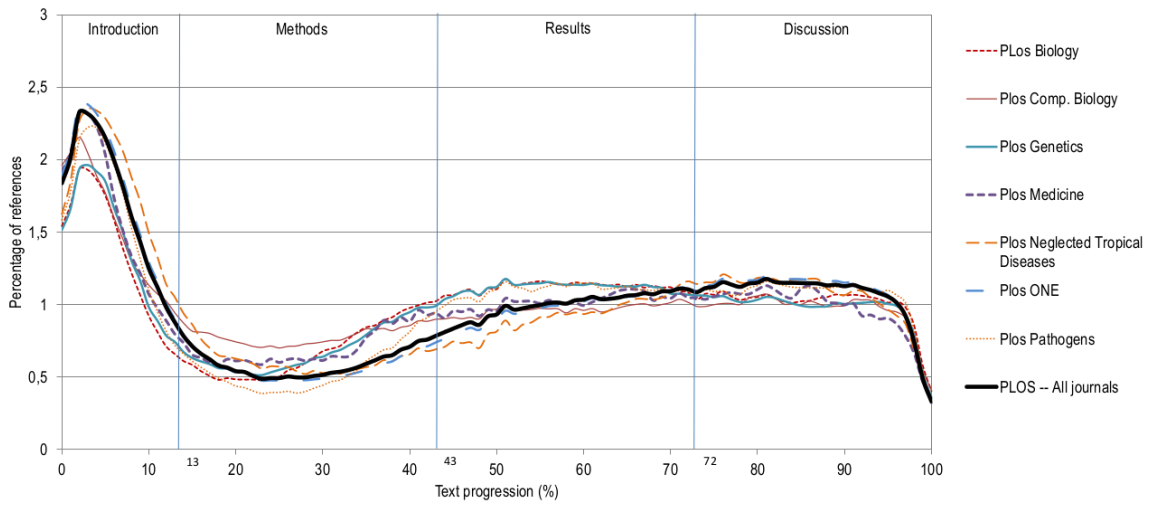


Figure : Distribution of References of PLOS Journals

Research Problem

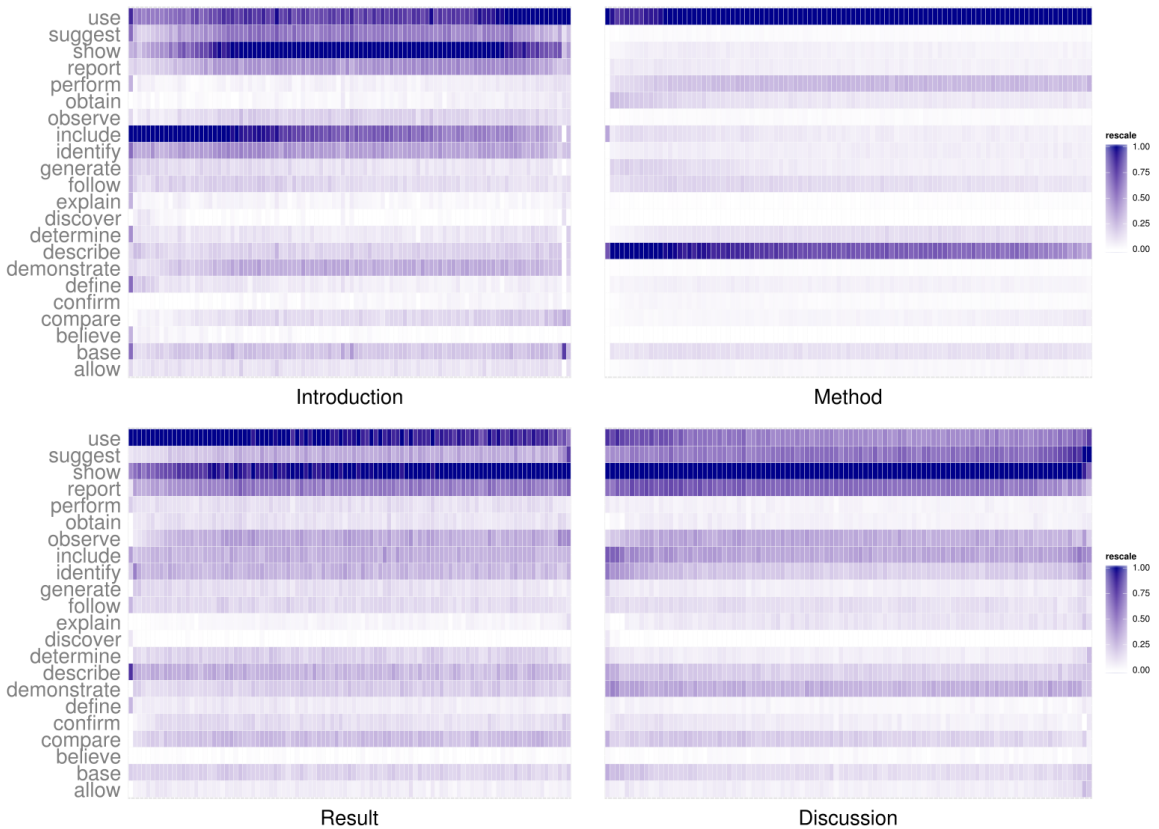


Figure : Density of Verbs in Citation Context

Method

- Correspondence Factorial Analysis (CFA) - a multivariate statistical method. Correspondence analysis is a technical description of contingency tables and is mainly used in the field of text mining.
- Analysis of a dataset of about 48,000 textual contexts of bibliographical references (in-text citations).

Dataset

Journal	Articles	Citations	Citation contexts
PLOS Biology	1,754	170,785	91,117
PLOS Computational Biology	2,560	243,488	126,870
PLOS Genetics	3,414	332,845	185,5 37
PLOS Medicine	926	72,676	34,819
PLOS Negl. Tropical Diseases	1,872	133,022	73,211
PLOS ONE	72,158	5,363,036	2,854,082
Total	*82,684*	*6,315,852*	*3,365,636*

- Published by the Public Library of Science (PLOS), in Open Access
- XML, Journal Article Tag Suite (JATS)
- Entire corpus up to September 2013

Processing

Processing steps:

- Identification of the section structure of articles by analyzing the section titles.
- Sentence segmentation.
- Extraction of all textual segments (sentences) that contain in-text citations in each section type: we obtain a total of more than 3 million sentences in 4 different sets corresponding to each section.

Corpus subset

Corpus subset of sentences:

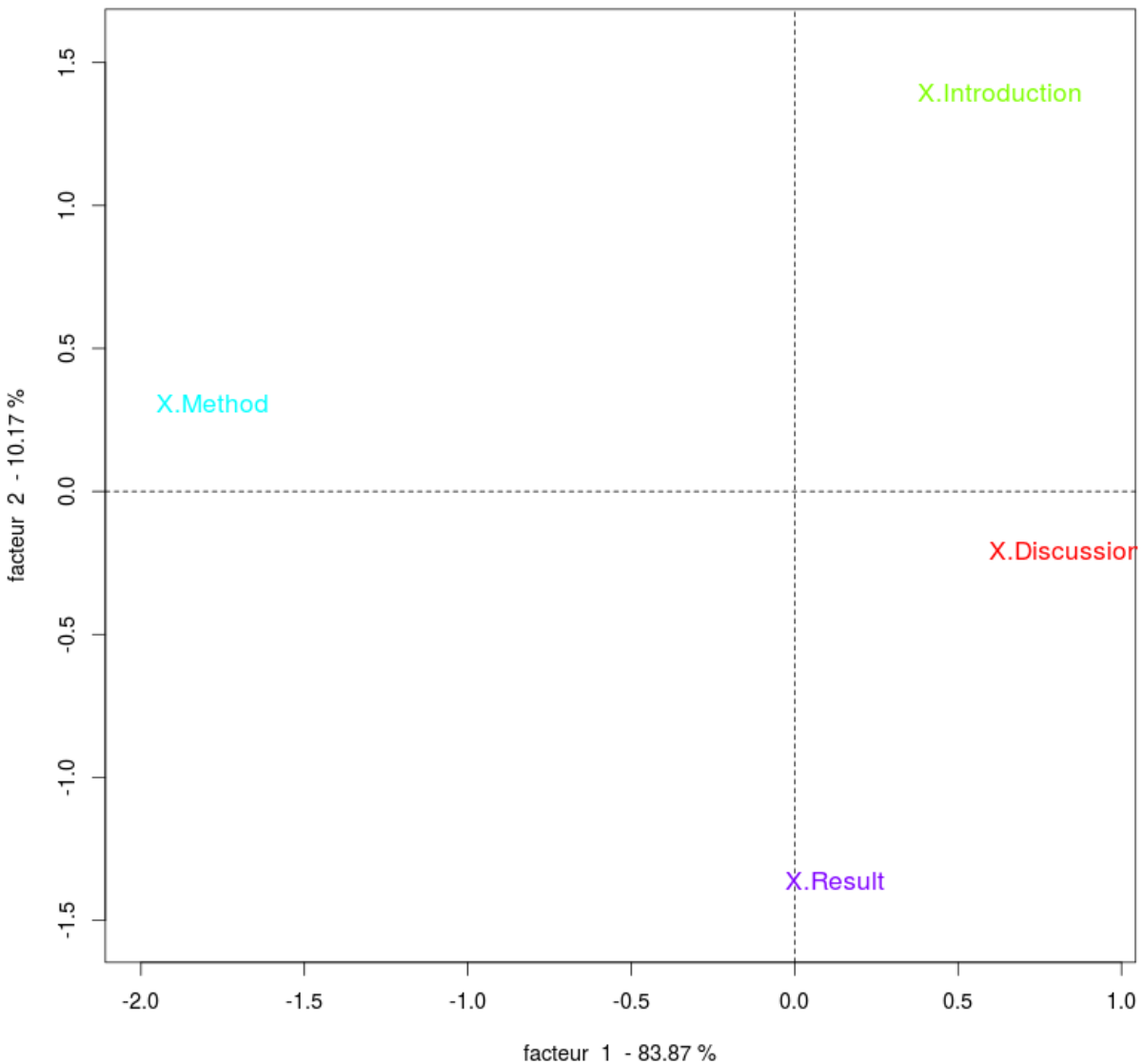
- We take 2,000 randomly extracted sentences for each section of the rhetorical structure and for each journal: a total of 48,000 sentences.
- The subset contains 47,714 unique terms, that have 1,569,201 occurrences (see table below).

Vocabulary Summary by Section Type

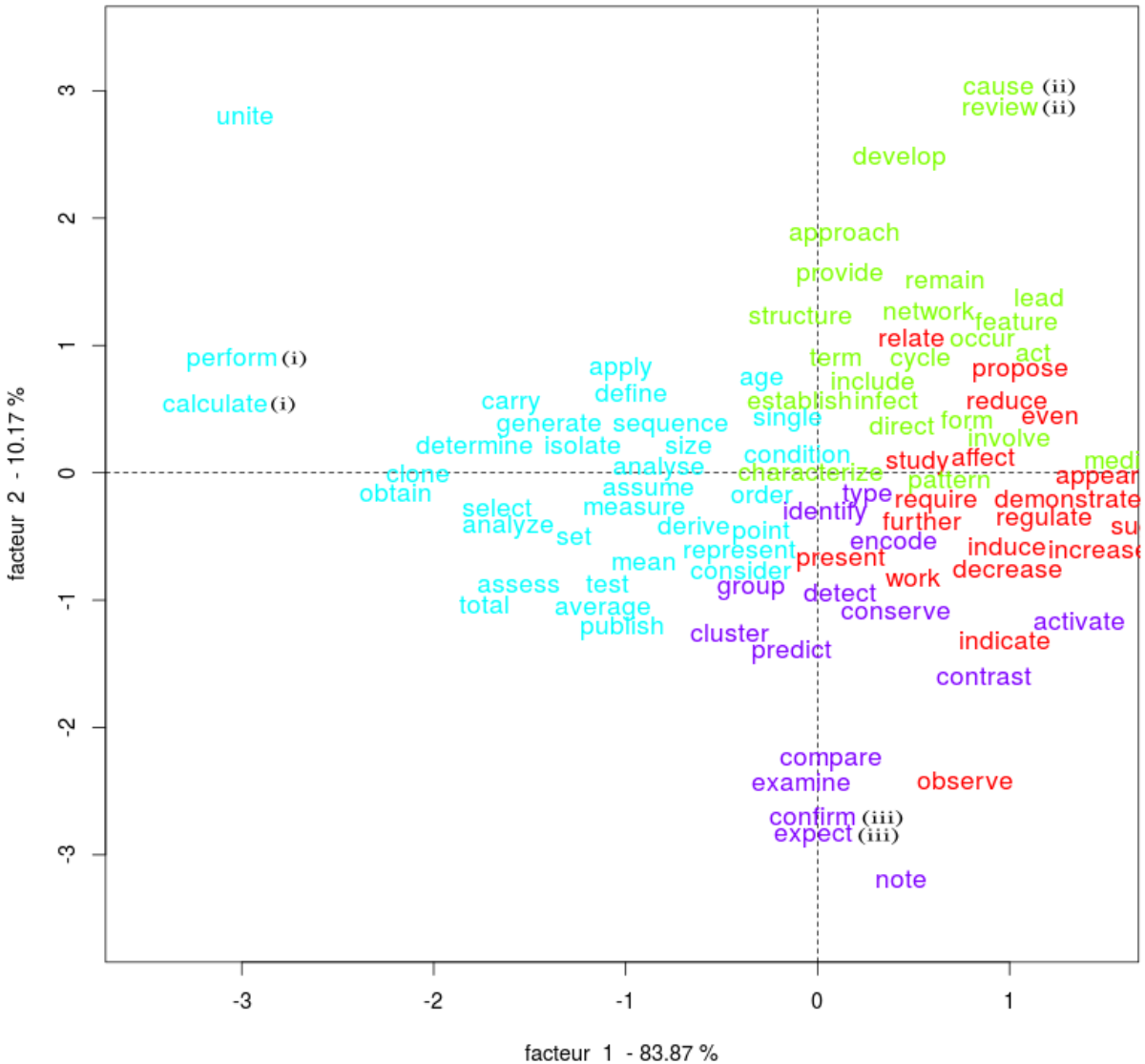
	I	M	R	D	*Total*
Nb terms	327,506	295,798	337,379	608,518	*1,569,201*
Nb unique terms	21,389	22,848	22,316	28,694	*47,714*
% unique terms	6.5	7.7	6.6	4.7	*3.0*
Nb hapax legomena	8,809	10,539	9,326	11,689	*18,082*
% hapax legomena	2.7	3.6	2.8	1.9	*1.2*
Nb words	327,506	295,798	337,379	608,518	*1,569,201*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

Processing of corpus subset

- Tokenization and lemmatization, stemming from dictionaries, without disambiguation.
- Filter all verb forms and calculate occurrence frequencies for each section type.
- Factorial Correspondence Analysis using the occurrence frequency tables for the most frequent verbs.



CA - Projections of Sections on a Factorial Plane



CA - Projections of Most Frequent Verbs on a Factorial Plane

Relative Frequency of Verbs

Verbs	Discussion	Introduction	Methods	Results
...	...	...	...	...
calculate (i)	1.32	1.18	**22.47**	3.89
cause (ii)	9.35	**16.79**	3.64	5.08
confirm (iii)	5.74	2.96	5.11	**9.01**
...	...	...	...	...
expect (iii)	5.57	2.96	3.93	**7.96**
...	...	...	...	...
perform (i)	3.87	3.82	**52.16**	7.87
...	...	...	...	...
review (ii)	7.95	**14.06**	2.61	4.27
...	...	...	...	...

Conclusions (1)

- We have shown that citation contexts are strongly dependent on the rhetorical structure.
- By considering the most frequent verbs in the different sections, our results imply functionality contexts which are specific to the rhetorical structure.
- Therefore, for the analysis of citation acts, it is necessary to take into consideration the rhetorical structure of articles.

Conclusions (2)

- 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.
- 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.

Thank you for your attention!

Marc Bertin
Post-doctoral Fellow
Université de Québec à Montréal, Canada
bertin.marc@gmail.com

Iana Atanassova
Assistant Professor
Centre Tesniere, University of Franche-Comte, France
iana.atanassova@univ-fcomte.fr

Related work

Marc Bertin and Iana Atanassova.
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)*, pages 5–12, Amsterdam, The Netherlands, April 13 2014.

Marc Bertin, Iana Atanassova, Vincent Larivière, and Yves Gingras.
The distribution of references in scientific papers: an analysis of the imrad structure.
In *14th International Society of Scientometrics and Informatics Conference*, Vienna, Austria, 15-19th July 2013. International Society for Scientometrics and Infometrics.

Marc Bertin, Iana Atanassova, Vincent Larivière, and Yves Gingras.
The invariant distribution of references in scientific papers.
Journal of the Association for Information Science and Technology (JASIST), 2014 (in press).

Marc Bertin, Iana Atanassova, Vincent Larivière, and Yves Gingras.
The linguistic context of citations: a cartography of the structure of scientific papers.
In *American Association for the Advancement of Science (AAAS) Annual Meeting*, San Jose, California, 2015.

Marc Bertin, Iana Atanassova, Vincent Larivière, and Yves Gingras.
Mapping the Linguistic Context of Citations.
Bulletin of the Association for Information Science and Technology (ASIST) Featuring the "The Future of Science Mapping", 41(2), December/January 2015.