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

Marc Bertin and Iana Atanassova

April 13, 2014
CIRST - UQAM, Montreal, Canada

Bibliometric-enhanced Information Retrieval
ECIR-2014, Amsterdam

Research Problem

Scientific papers usually follow a specific rhetorical structure:

IMRaD (Introduction, Method, Result and Discussion)

Questions

- What relationships exist between citation contexts and the structure of the papers?
- How does the IMRaD structure affect the occurrences of verbs in citation contexts?

Hypothesis

Verbs that appear close to bibliographic citations in texts most frequently define the relation between the article's author and the cited work.

Dataset

Journal	Articles	Citations	Citation contexts
PloS Biology	1,587	150,429	79,703
PloS Computational Biology	1,976	177,742	92,437
PloS Genetics	2,435	227,121	126,230
PloS Negl. Tropical Diseases	1,240	83,402	45,714
PloS Pathogens	2,208	209,685	115,750
Total	*9,446*	*848,379*	*459,834*

- Published by the Public Library of Science (PLoS)
- XML, Journal Article Tag Suite (JATS)
- Entire corpus up to September/October 2012

Method

1. Identification of the section structure of the articles
2. Sentence segmentation and citation contexts extraction
3. POS-tagging & lemmatization to identify verbs in citation contexts
4. Comparisons of the ranked verb lists for the sections

Method

1. Identification of the section structure of the articles
   - Section titles were analysed in order to match each section with one of the section types in the IMRaD structure.
2. Sentence segmentation and citation contexts extraction
3. POS-tagging & lemmatization to identify verbs in citation contexts
4. Comparisons of the ranked verb lists for the sections

Method

1. Identification of the section structure of the articles
   - Section titles were analysed in order to match each section with one of the section types in the IMRaD structure.
2. Sentence segmentation and citation contexts extraction
   - We detect sentence boundaries by analyzing the punctuation using a set of typographic rules.
   - Sentences are basic linguistic units of text, suitable to model text progression.
   - Sentence boundaries as delimiters of citation contexts.
3. POS-tagging & lemmatization to identify verbs in citation contexts
4. Comparisons of the ranked verb lists for the sections

Method

1. Identification of the section structure of the articles
   - Section titles were analysed in order to match each section with one of the section types in the IMRaD structure.
2. Sentence segmentation and citation contexts extraction
   - We detect sentence boundaries by analyzing the punctuation using a set of typographic rules.
   - Sentences are basic linguistic units of text, suitable to model text progression.
   - Sentence boundaries as delimiters of citation contexts.
3. POS-tagging & lemmatization to identify verbs in citation contexts
   - Stanford TreeTagger
4. Comparisons of the ranked verb lists for the sections

Results

- Processed $\sim$ 460,000 citation contexts.
- 1807 verbs have occurrences in all four sections.

Distribution of verbs in sections (Zipf's law)

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

Top 10 verbs per section

Rank	Introduction	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

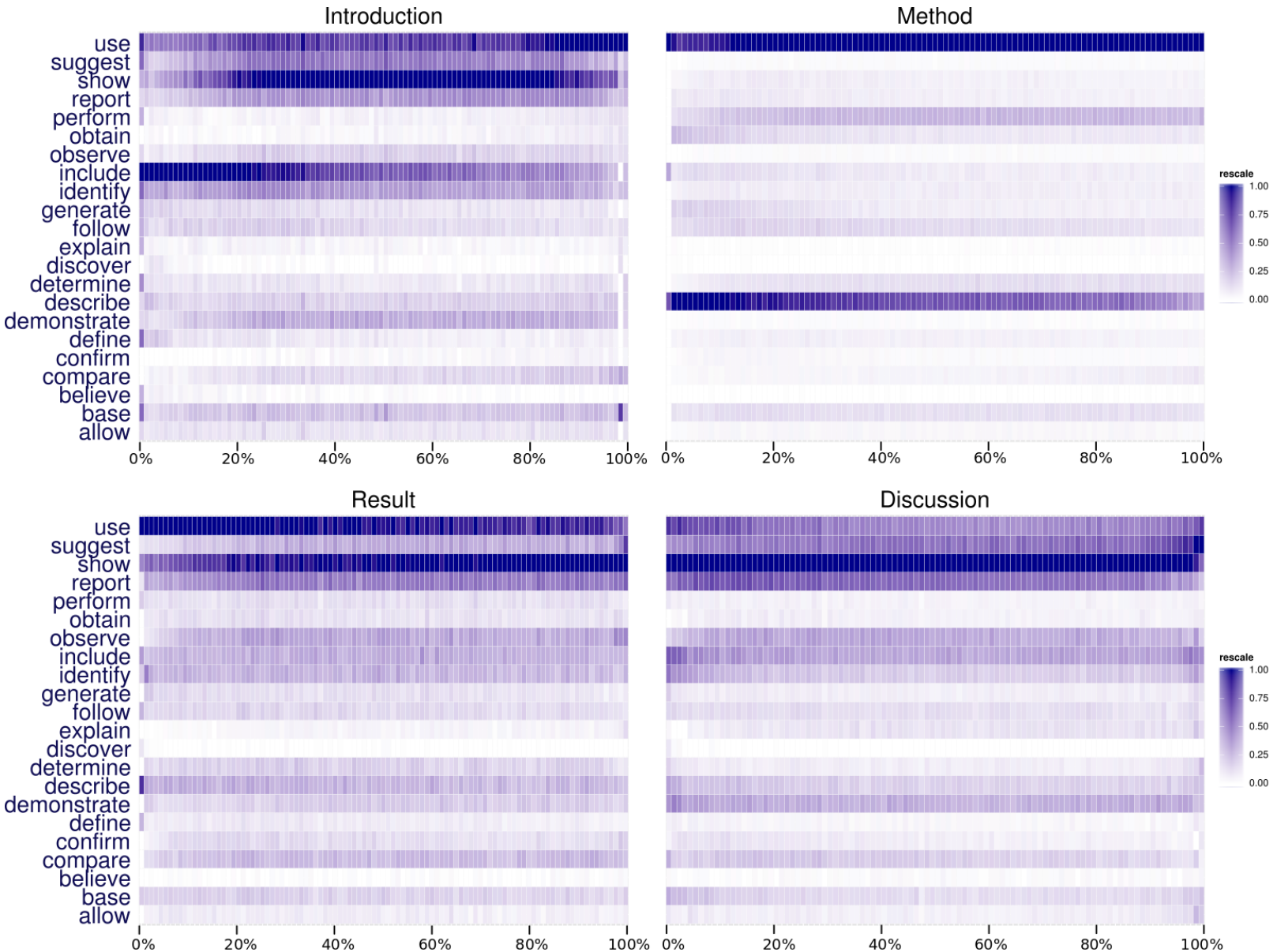
Top 10 verbs per section

Rank	Introduction	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

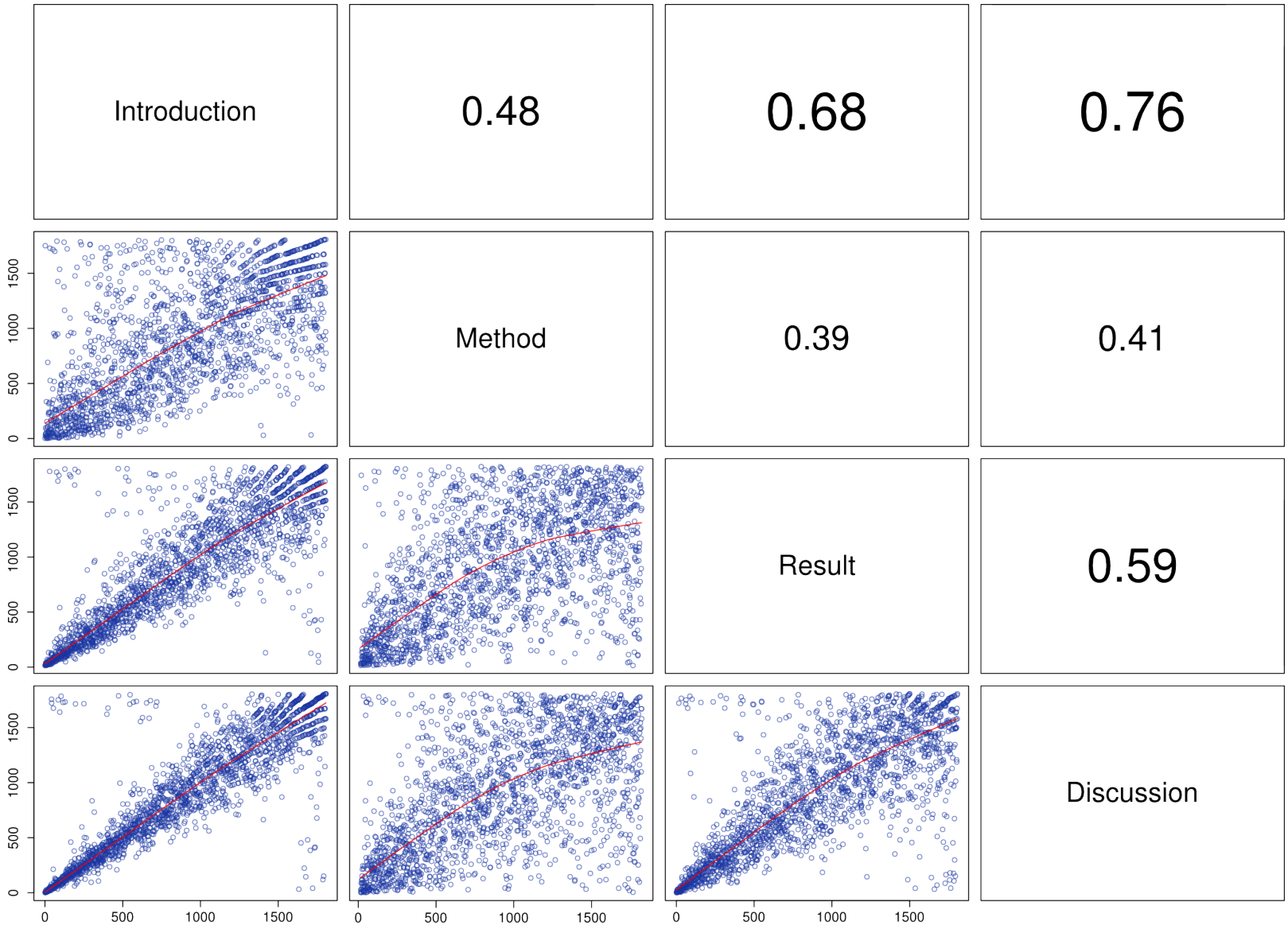
Top 10 verbs per section

Rank	Introduction	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

Densities of verbs in citation contexts



Scatterplots of section pairs and values for Kendall τ



Conclusions

- Citations play different roles according to their position in the rhetorical structure.
- Citation acts are expressed by a small number of verbs.
- The study of citation act verbs is a step towards the categorization of citations and network structures.

Thank you for your attention!

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

Iana Atanassova
Post-doctoral Fellow
Concordia University, Montreal, Canada
iana.atanassova@gmail.com

Related work

Iana Atanassova and Marc Bertin.
Faceted Semantic Search for Scientific Papers.
In *The Semantic Publishing Challenge, Task 3: In-use task, 11th European Semantic Web Conference (ESWC-2014)*, Crete, Greece, May 2014.

Marc Bertin and Iana Atanassova.
Semantic Enrichment of Scientific Publications and Metadata.
D-Lib Magazine, 18(7/8), 2012.

Marc Bertin and Iana Atanassova.
Hybrid Approach for Semantic Processing of Scientific Papers.
In *The Semantic Publishing Challenge, Task 2: Extraction and characterization of citations, 11th European Semantic Web Conference (ESWC-2014)*, Crete, Greece, May 2014.

Marc Bertin, Iana Atanassova, Vincent Larivière, and Yves Gingras.
The Distribution of References in Scientific Papers: an Analysis of the IMRaD Structure.
In *Proceeding of 14th International Society of Scientometrics and Informetrics Conference*, Vienna, Austria, 15th-19th July 2013. International Society for Informetrics and Scientometrics.

Marc Bertin, Iana Atanassova, Vincent Larivière, and Yves Gingras.
The Cognitive Context of Citations.
In *10th Iteration of the Places & Spaces: Mapping Science Exhibit on "The Future of Science Mapping"*, 2014.