The Invariant Distribution of References in Scientific Articles

Marc Bertin[1], Iana Atanassova[1], Vincent Larivière[2] and Yves Gingras[3]

[1] bertin.marc@courrier.uqam.ca, iana.atanassova@nlp-labs.org

Observatoire des Sciences et des Technologies (OST), Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Québec à Montréal, CP 8888, Succ. Centre-Ville, Montréal, QC. H3C 3P8 (Canada)

[2]vincent.lariviere@umontreal.ca

École de bibliothéconomie et des sciences de l'information, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, QC. H3C 3J7 (Canada) and Observatoire des Sciences et des Technologies (OST), Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Québec à Montréal, CP 8888, Succ. Centre-Ville, Montréal, QC. H3C 3P8 (Canada)

[3]gingras.yves@uqam.ca

Observatoire des Sciences et des Technologies (OST), Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Québec à Montréal, CP 8888, Succ. Centre-Ville, Montréal, QC. H3C 3P8 (Canada)

Abstract

The organization of scientific papers typically follows a standardized pattern, the well-known IMRaD structure (Introduction, Methods, Results and Discussion). Using the full-text of 45,000 papers published in the PLOS series of journals as a case study, this paper investigates, from the viewpoint of bibliometrics, how references are distributed along the structure of scientific papers, as well as the age of these cited references. It shows that, once the sections of articles are realigned to follow the IMRaD sequence, the position of cited references along the text of articles is invariant across al PLOS journals, with the introduction and conclusion accounting for most of the references. It also provides evidence that the age of cited references varies quite a lot by section, with older references being found in the methods and more recent ones in the discussion. On the whole, these results provide insights into the different roles citations have in the scholarly communication process.

Introduction

In recent years, the full-text access to scientific papers in machine-readable formats has provided possibilities for the development of novel and comprehensive approaches for the analysis of citations, argumentative structures and the distribution of references (Ding et al., 2013, Liu & Chen, 2013, Bertin et al., 2013, Small, 2011, Teufel, 2006, Bertin & Atanassova, 2012). The distribution of references in the body of scientific papers can be studied on a large scale in order to obtain detailed descriptions of these phenomena and observe new trends.

The interest in the localization of references in papers is far from recent, however, and can be traced back to the 1970s. Voos and Dagaev (1976) raised the question whether all citations should be considered

equally in citation analysis. They conducted a manual study of the distribution of references by section for four papers. Although the small size of their sample did not allow to any generalization, they showed the subject to be of interest for further research: *"In terms of sampling theory this is certainly not a sufficient number to provide a definitive answer to the questions of value being posed, but it is sufficient to demonstrate that is a valid area for further research"* (Voos & Dagaev, 1976, p.19).

Citation location was used by Cano (1989) as a variable for the characterization of citation behavior in the model of Moravcsik and Murugesan (1975). To characterize the function of references in texts, these studies have focused on three main variables: the position of citations in articles, the context of citations, and the age of the references. Cano proposed an empirical study of 344 citations to examine the *"usefulness of a citation location parameter as a novel bibliometric variable to refine the crudeness of pure citation counts".* He calculated the normalized distance of a citation from the beginning of the article by measuring its coordinates within printed pages. His study identified three areas of citation concentration: beginning section (up to the 15th percentile), middle section (from the 20th to the 75th percentile) and end (from the 80th percentile). Furthermore, Cano showed that the largest concentration of references is in the first 15% of a paper, which is consistent with the previous results reported by Voos and Dagaev (1976).

The idea that the section structure of papers plays an important role in determining the function and importance of citations was first developed by McCain and Turner (1989). They analyzed citation locations using the sections in the citing paper and proposed a method to weigh citation occurrences in order to construct a utility index of papers. To do this, they used several assumptions related to the nature of the sections, assigning different weights according to their rhetorical functions in a citation context classification scheme.

Later, Maričić et al. (1998) studied a collection of 357 papers focusing on three components: locations of references, levels of citation, and age. They suggested that if the section structure is derived from publishing practices, it also reflects the structure of scientific articles. As a result, references have different values according to their location, i.e. the section in which they appear. To express these differences they assigned weights to the different sections using a ranking scale (Introduction: 10, Methods:30, Results:30, Discussion:25).

More recently, several studies have focused on the localization and distribution of references by examining the structure of articles. By using bibliometric methods and considering the position of sections Hu et al. (2013) analyzed 350 papers published in the *Journal of Informetrics*. They counted the citations per section and showed that the first sections have a higher density of citations than the following sections. Ding et al. (2013) performed an analysis of cited works in 866 articles from the *Journal of the American Society of Information Science and Technology*. They studied the number of times each reference was cited across sections and obtained citation frequencies per section. Both Hu et al. (2013) and Ding et al. (2013) operated on the section level and did not record positions of citations within sections, at the level of sentences.

Other related variables have also been considered to characterize the function of citations. For example, Wan and Liu (2014) proposed to measure the importance of citations by using six features: number of occurrences, location in sections, time interval, average length of citing sentences (number of words in sentence), average density of citation occurrences and self-citations. By experimenting with 40 papers in computer science, they showed, using a regression method, that good correlation can be achieved with human-assigned values.

This paper contributes to this literature by studying references found in scientific papers in relation to the categories of the IMRaD structure of scientific papers (Introduction, Methods, Results and Discussion). The IMRaD structure is intended to facilitate reading and access to information but, more precisely, it can be viewed as a standardization of scientific writing providing a framework for argumentative logic. This structure was considered as an ideal in the early twentieth century and has imposed itself in most major scientific journals in the mid-twentieth century, becoming the main standard in the 1970s. For example, the field of Physics has adopted it extensively in the 1950s, and in the 1980s it became predominant also in Medicine (Sollaci & Pereira, 2004). Many studies have focused on various aspects of this structure: automatic classification of sentences in full-text (Agarwal and Yu, 2009), the effects of the use of the IMRaD style (Oriokot et al., 2011), creation of guidelines for scientific writing (Kucer, 1985; Meadows, 1985; Day and Gastel, 2006), production of structured abstracts (Nakayama et al., 2005) and editorial requirements (Barron, 2006).

In this study, we examine the distribution of references along the four sections of the IMRaD structure, working at the sentence level. This allows us to obtain the distribution of references along the progression of sections and identify high and low density zones within this argumentative structure. Sentence-level processing is relatively recent in such studies (Wan and Liu, 2014), but it has been already used in other domains. For example, Agarwal and Yu (2007) have proposed to automatically classify sentences in the IMRaD structure. The applications of such approaches are mainly in the field of Information Retrieval and question answering system. However, the interest of sentence-level processing is that, from a linguistic viewpoint, sentences are minimal units having a semantic meaning that tends to be relatively contextually independent.

This paper aims at answering the following research questions. First, how are references distributed within the text and its different sections? Second, is there a difference in the age of references depending on where they appear in the document? To study the distribution of reference along the IMRaD structure, we characterize references along two criteria: their position along the text and the age of the cited documents. Our motivation is to provide evidence of the existence of a relation between the distribution of references and the argumentative structure of scientific writing. Previous studies showed that the distribution of references is heterogeneous. However, this phenomenon needs to be studied more precisely in order to determine if, taking into account the variations in the order of presentations, one could not find an invariant distribution.

This analysis is performed using the entire PLOS corpus up to September/October 2012, which contains about 45,000 articles and 197,000 sections. The automatic processing of the papers makes possible working with much bigger document collections compared to datasets used in previous studies conducted manually (up to 500 papers). The necessity of such large-scale studies lies in the fact that working with bigger datasets provides insight into global phenomena and the results are less likely to be biased by the singular properties of any particular document present in the collection.

Methods

Our analysis proceeds in several steps: a) categorisation of the sections of the text according to section titles; b) segmentation of the text into sentences in order to obtain the distribution of the references along the text; c) identification of the in-text references; d) reconstruction of the IMRaD structure.

To avoid confusion, we recall that we will use the term “reference” to denote items in the bibliography of a

paper, as defined by Derek de Solla Price. We use the term “in-text reference” when we refer to references that appear in the text of a paper and that point to items in the bibliography. By definition, the number of in-text references in a paper is superior or equal to the number of references in the bibliography.

Data source

We performed an automatic analysis of a large document collection of articles published in the PLOS series of journals. Founded in 2006, the Public Library of Science (PLOS) is an Open Access publisher of seven peer-reviewed academic journals. PLOS ONE, the publishers’ general journal covers all fields of science and social sciences, divided into eleven subject areas: Biology and life sciences, Computer and information sciences, Earth sciences, Ecology and environmental sciences, Engineering and technology, Medicine and health sciences, People and places, Physical sciences, Research and analysis methods, Science policy, and Social sciences. This makes the PLOS corpus a prime candidate to analyze the properties of the distribution of references throughout articles in relation to the IMRaD sequence, independently of the discipline.

As these seven journals follow the same publication model but are in different scientific fields, our aim is to observe the different uses of bibliographic references in these fields and their relation to the structure of the articles. PLOS provides access to the articles in XML format. The set of XML elements and attributes that are used for the representation of journal articles are known as Journal Article Tag Suite (JATS), which is an application of Z39.96-2012 (ANSI 2012) and a continuation of the NLM Archiving and Interchange DTD work by NCBI (http://dtd.nlm.nih.gov/). Some studies (Carter, Funk & Mooney 2012) give various applications of this standard. The JATS structure of an article consists of three main elements *front* – *body* – *back*, where the textual content of the article is in the body element, which is further divided into sections and paragraphs. The *<front>* tag contains some traditional fields of metadata (title, authors, etc.) as well as the article type, and the *<back>* tag contains, among others, the bibliography.

The availability of the corpus in XML format is an important factor for automatic text processing because it gives access to the articles’ content in structured full text. This makes possible the development of fine-grained full-text processing for studying the properties of the different sections of the articles. Other publishers often give access to scientific articles only in PDF format, which implies heavy pre-processing and error-prone text extraction in order to obtain the text content of the articles and detect the section structure. The articles in the PLOS journals are categorized in several different types, such as “Research article”, “Perspective”, “Review”, etc. Table 1 presents the different article types in the PLOS corpus, exploiting the metadata present in the XML documents. The article types are identified using the contents of the *<article-meta>* tag in the JATS structure.

Table 1: Article Types in PLOS Journals

Document type	PLOS Biology	PLOS Comp. Biology	PLOS Genetics	PLOS Medicine	PLOS N. T. Diseases	PLOS Pathogens	PLOS ONE	Percentage
Research Article	1 552	1 876	2 373	782	1 154	2 142	33 721	92,06%
Synopsis	775			128				1,91%
Perspective	35	49	63	260				0,86%
Correspondence	9	14	5	255				0,60%
Review		32	56	16	48	59	38	0,53%
Editorial	40	39	9	91	50	6		0,50%
Essay	75			139				0,45%
Policy Forum				207				0,44%
Correction	59	23	16	57	1	16		0,36%
Primer	156							0,33%
Opinion						81		0,17%
Health in Action				77				0,16%
Community Page	69							0,15%
Book Review/Science in the Media	68							0,14%
Feature	53							0,11%
Pearls						50		0,11%
Research in Translation				49				0,10%
Viewpoints			6		42			0,10%
Other article types	74	74	32	167	71		23	0,93%
Total	*2 965*	*2 107*	*2 560*	*2 228*	*1 366*	*2 354*	*33 782*	*100,00%*

This table shows, as could be expected, that the 'Research Article' type is largely dominant and accounts for more than 92% of all articles in the corpus. Moreover, 99.82% of the PLOS ONE articles are of the 'Research Article' type. Among the other six journals, PLOS Biology and PLOS Medicine stand out as having a relatively small number of Research Articles: 52.34% for PLOS Biology and 35.10% for PLOS Medicine. In fact, these two journals offer a wider variety of article types like 'Synopsis', 'Perspective', 'Correspondence', 'Essay', 'Policy Forum' and 'Primer'. This phenomenon needs further investigation and suggests disciplinary differences in the use of these document types. The question arises whether this is related to the nature of the discipline or is rather the result of an editorial choice.

Table 2 presents the number of articles processed for each journal, as well as the average number of sections and sentences per article. It shows that the average number of sections per article varies between 3.48 and 4.74 according to the journal. The two journals PLOS Medicine and PLOS Biology have the smallest number of sections and this result is related to the fact that these journals are composed of various article types, as mentioned above. We can also observe that the average length of articles varies significantly across journals: from 96 sentences for PLOS Medicine, to 242 sentences for PLOS Computational Biology. The Table also shows the relative importance of PLOS ONE: papers published in this journal account for more than 71% of all the papers in the corpus.

Table 2. Descriptive Statistics on PLOS Journals

Journal	Number of articles	Avg number of sections	Avg number of paragraphs	Avg number of sentences	Avg number of in-text references	Avg number of references
PLOS Biology	2 965	3,48	26,17	141,77	54,63	33,99
PLOS Comp. Biology	2 107	4,69	46,91	242,00	87,49	52,25
PLOS Genetics	2 560	4,74	39,33	218,80	91,09	55,66
PLOS Medicine	2 228	4,10	21,61	95,98	39,62	28,27
PLOS Negl. Trop. Diseases	1 366	4,50	30,46	157,44	67,57	43,70
PLOS Pathogens	2 354	4,74	36,30	216,91	93,41	56,98
PLOS ONE	33 782	4,47	33,29	177,90	74,10	47,28
All PLOS journals	*47 362*	*4,43*	*33,30*	*178,19*	*73,55*	*46,61*

For the following processing stages we will be mostly interested in the "Research article" type. However it

is interesting to observe the differences between the different document types. Table 3 presents statistics on the eight most frequent document types. It shows clearly that research articles have characteristics similar to some of the other article types such as “Review”, “Essay” and “Policy Forum”, that tend to have more than five sections and between 88 and 188 sentences on average. Other article types, such are “Synopsis”, “Perspective”, “Correspondence” and “Editorial” are, unsurprisingly, much shorter and contain a smaller number of sections.

Table 3: Descriptive Statistics of Document Types Published by PLOS Journals

Document type	Number of articles	Avg number of sections	Avg number of paragraphs	Avg number of sentences	Avg number of in-text references	Avg number of references
Research Article	43 600	4,53	34,88	187,94	77,76	48,94
Synopsis	903	1,13	6,86	27,27	0,00	0,01
Perspective	407	4,06	12,67	57,09	19,86	16,74
Correspondence	283	1,06	4,75	19,22	5,79	4,77
Review	249	6,47	36,29	188,26	128,28	91,83
Editorial	235	2,41	10,63	45,83	15,27	11,01
Essay	214	5,35	20,02	88,30	34,83	31,85
Policy Forum	207	6,29	26,52	103,62	42,52	36,31

Our processing follows several stages that will be described in detail below. First, the *body* elements of the XML documents were processed and the sections were categorized, to verify the coherence of the corpus with the IMRaD structure. We then processed the text content of all paragraphs in order to segment them into sentences. This segmentation allows us to work with text elements that are smaller than paragraphs so that we can associate the bibliographic references with given sentences and obtain their distribution along the text. Finally, our algorithm counts the number of references in each sentence.

Identification of the structure of the text

As our aim is to characterize the IMRaD structure, we need to identify in our corpus the sections that belong to the four main types: Introduction, Methods, Results and Discussion. Articles that follow the IMRaD structure can in fact contain more than four sections (e.g. having an additional fifth section named “Supplementary Material”), or can contain the four IMRaD sections but in a different order. For this reason, the section position is not a sufficient criterion to identify the type of a section. In our case, to perform this analysis on a large dataset (197,335 sections), we need to implement an automatic section categorization algorithm.

In general, each section of a paper has a title and a content that consists of paragraphs and non-textual block elements such as tables and figures. We have analyzed all section titles that are present as separate elements in the XML documents. We suppose that section titles contain the most reliable indication on the type of the section, as section titles express the general argumentative structure of an article. However, the title that expresses a given section type can vary from one article to another. For example, a section corresponding to the “Method” in the IMRaD structure can be signalled by several different titles such as “Method”, “Methods”, “Method and Model”, etc. In fact, while the structure itself remains fixed, many variations are possible in the formulation of the section titles and these must be taken into consideration during the processing. To overcome these problems, we have constructed sets of regular expressions that match each of the four types of the IMRaD structure.

The number of categorized sections in all research articles is presented on Table 4. The columns “Methods and Results” and “Results and Discussion” account for the cases where two different sections of the IMRaD

structure are combined into a single one.

Table 4: Section categories in research articles

Journal	Introduction	Methods	Methods and Results	Results	Results and Discussion	Discussion	Supporting Information	Other sections
PLOS Biology	1 550	1 545	2	1 346	202	1 344	1 357	26
PLOS Comp. Biology	1 876	1 787	0	1 636	240	1 632	1 510	122
PLOS Genetics	2 373	2 366	0	2 104	268	2 104	2 210	18
PLOS Medicine	780	778	0	776	2	776	555	22
PLOS Negl. Trop. Diseases	1 153	1 155	0	1 083	70	1 082	701	14
PLOS Pathogens	2 142	2 142	0	2 005	2 891	2 003	1 782	4
PLOS ONE	33 717	33 463	9	30 657	137	30 760	18 613	445
All PLOS journals	*43 591*	*43 236*	*11*	*39 607*	*3 810*	*39 701*	*26 728*	*651*

Table 4 shows that almost all sections were categorized either as one of the IMRaD types or as "Supporting information". The remaining 651 sections could not be categorized because their titles did not correspond to any of the IMRaD types. Among those, some of the most frequent titles are: "Author Summary", "Accession Numbers", "Analysis" and "Acknowledgements".

Having identified the section types, we consider the articles that contain the four section types (Introduction, Methods, Results and Discussion). Then, we can study the positions of these sections and the distribution of references along this IMRaD structure.

Table 5 presents the number of articles in each journal that contain the four section types and their percentage of all research articles. The papers in which two of the sections are combined into a single one are also included in this table provided that the remaining two sections in the IMRaD structure are also present.

Table 5: Number and percentage of research articles that contain the four section types of the IMRaD structure

Journal	Research articles	Articles that contain all four IMRaD types	Percentage
PLOS Biology	1 552	1 539	99,16%
PLOS Comp. Biology	1 876	1 785	95,15%
PLOS Genetics	2 373	2 365	99,66%
PLOS Medicine	782	778	99,49%
PLOS Negl. Trop. Diseases	1 154	1 152	99,83%
PLOS Pathogens	2 142	2 140	99,91%
PLOS ONE	33 721	33 384	99,00%
All PLOS journals	*43 600*	*43 143*	*98,95%*

This table shows that almost 99% of all research articles in the corpus contain the four section types of the IMRaD structure. All seven journals in our study systematically use the IMRaD structure. The lowest value is for PLOS Computational Biology, 95.15%.

If we consider all document types in the corpus (see Table 1), the percentage of research articles that contain all four IMRaD types is 91.1%. This result can be compared with a previous result presented by (Bertin et al. 2013) for the same corpus, where it was shown that about 83% of the articles contain the four sections that are typically found in the IMRaD structure. Here, this percentage is slightly higher due to the fact that we consider also the cases where two section types are combined into a single section as

explained above.

Sentence level processing

Since we are interested mainly in the text content of the documents, the further processing stages consider only the text present in the paragraph elements in each section, thus eliminating all tables and figures.

The JATS structure used by PLOS provides paragraph elements as the finest level of text segments. For our analysis, we needed segmentation into sentences and we parsed the initial JATS trees in order to extract the relevant text segments from the article body, as well as other elements such as sections, section titles, section numbers, paragraphs and the bibliography. These data were stored in the DocBook format that was used as the basis for the further processing.

Each paragraph was segmented into sentences by analysing the punctuation of the text following a set of typographic rules. All the occurrences of symbols denoting sentence boundaries (point, exclamation mark, etc.) were examined and disambiguated. The occurrence of a point in a text does not necessarily mean a sentence end, because in many cases it can be part of an abbreviation, references, genus species, numeric values, etc. We used a set of finite-state automata in order to determine the contexts in which the points signal sentence ends. For this purpose, we have developed a Java application based on the work of (Mourad 2001). The algorithm uses a rule-based approach which disambiguates the use of punctuation marks by examining the close context of their occurrences. This analysis showed that more than 60% of all occurrences of points in the corpus do not correspond to the end of a sentence. Once we have identified the sentence boundaries in the corpus, we can consider the sentences as the finest textual unit and examine the number of references in each sentence. In fact, a sentence can contain one or more references or an enumeration of references, which is rather frequent in the background section or the introduction.

In-text reference processing

Once we have identified sentence boundaries, we examine each sentence and count the number of in-text references. As the input data is in the XML JATS format, most of the in-text references are marked with <*xref*> tags, that link the reference with an item of the bibliography at the end of the article. However, simply counting these tags is not a reliable method to obtain the reference counts. In fact, the XML structure does not render all of the actual in-text references, because of some specific typographic rules. In particular, the cases of reference range, where multiple sources are cited but only the first and the last sources are present as *xref* elements (like “6-9”, are rather frequent in the corpus (on average more than one per article). Such cases must be taken into consideration and need additional processing. We have measured the importance of this phenomenon and found that in-text references that do not appear as elements in the XML markup account for about 8% of all in-text references in the corpus. To tackle this problem, we used typographic analysis in order to identify reference range and establish the missing links with the bibliography items. Similar method for reference processing was presented by Bertin et al. (2013).

Finally, obtaining publication years of the cited works is straightforward, as this information is present in the bibliography items found in the back elements of the XML documents. This allows us to obtain the reference years associated to each sentence.

Results

First, we examine the order in which the four sections of the IMRaD structure are present in the articles. Table 6 presents the results of the categorization of the sections for the seven PLOS journals. We consider the actual position of each section in the text and compare it with its type, so that Introduction corresponds to section one, Method corresponds to section two, Result corresponds to section three and Discussion to section four. In Table 6, the column "Other" corresponds to sections that do not belong to any of the IMRaD types, for example sections having domain-specific or descriptive titles.

Table 6 shows that *PLOS Medicine* and *PLOS Neglected Tropical Diseases* essentially follow the IMRaD structure. The values on the diagonal of the matrix are very high, which means that virtually all the articles containing sections that were categorized follow the IMRaD standard. Hence, the first column, which corresponds to section one, never includes Method, Results and Discussion. This is coherent with the structure generally presented in the literature. For *PLOS ONE* and *PLOS Computational Biology*, we see that, while the first value presented on the diagonal is more than 92%, other values on the diagonal are considerably lower (close to 50% for *PLOS ONE* and 20% for *PLOS Computational Biology*), which indicates that the usual order of sections in IMRaD is in fact changed. The Method section, can be found not only in section 2 as expected with IMRaD, but also in section 4 usually reserved for Discussion. The standardization proposed for extraction of titles takes into account such variations. This inversion explains that the Results section often appears in Section 2 instead of 3, and that the methods are presented at the end of the article (Section 4). Of course these papers do not respect completely the standard IMRaD structure. They contain all the sections constituent of the IMRaD structure but in a different order and we can expect that this will affect the distributions of references. Finally, for *PLOS Genetics, PLOS Pathogens* and *PLOS Biology*, we note that the order of the sections and titles for these journals also differs from IMRaD with Methods coming last instead of Second and Discussion third instead of fourth as in the standard IMRaD structure.

Table 6. Relation between position of section and title of section for PLOS journals

PLOS journal	Section order	Introduction	Methods	Results	Discussion	Other	Total
PLOS Medicine	Section 1	**40,1%**	0,0%	0,0%	0,0%	59,9%	100,0%
	Section 2	1,3%	**48,5%**	0,3%	2,1%	47,8%	100,0%
	Section 3	0,1%	1,7%	**48,4%**	2,0%	47,8%	100,0%
	Section 4	0,1%	0,9%	0,5%	**52,1%**	46,4%	100,0%
PLOS Neglected Tropical Diseases	Section 1	**89,1%**	0,0%	0,0%	0,0%	10,9%	100,0%
	Section 2	0,4%	**89,8%**	0,2%	0,4%	9,2%	100,0%
	Section 3	0,0%	0,2%	**88,7%**	1,0%	10,0%	100,0%
	Section 4	0,0%	0,3%	0,3%	**86,1%**	13,3%	100,0%
PLOS ONE	Section 1	**99,9%**	0,0%	0,0%	0,0%	0,1%	100,0%
	Section 2	0,1%	47,8%	**51,5%**	0,1%	0,6%	100,0%
	Section 3	0,0%	6,0%	47,9%	**45,8%**	0,4%	100,0%
	Section 4	0,0%	**46,9%**	0,1%	46,8%	6,2%	100,0%
PLOS Computational Biology	Section 1	**92,7%**	0,0%	0,0%	0,0%	7,3%	100,0%
	Section 2	0,9%	19,0%	**69,7%**	0,2%	10,3%	100,0%
	Section 3	0,0%	10,6%	22,6%	**59,3%**	7,5%	100,0%
	Section 4	0,1%	**60,1%**	0,2%	21,7%	17,9%	100,0%
PLOS Genetics	Section 1	**94,8%**	0,0%	0,0%	0,0%	5,2%	100,0%
	Section 2	0,2%	4,2%	**91,9%**	0,0%	3,7%	100,0%
	Section 3	0,0%	10,4%	4,5%	**81,7%**	3,4%	100,0%
	Section 4	0,0%	**82,2%**	0,0%	3,9%	13,9%	100,0%
PLOS Pathogens	Section 1	**93,5%**	0,0%	0,0%	0,0%	6,5%	100,0%
	Section 2	0,0%	7,2%	**86,0%**	0,0%	6,7%	100,0%
	Section 3	0,0%	5,9%	7,3%	**80,7%**	6,2%	100,0%
	Section 4	0,0%	**81,3%**	0,0%	7,2%	11,6%	100,0%
PLOS Biology	Section 1	**52,9%**	0,0%	0,0%	0,0%	47,1%	100,0%
	Section 2	0,6%	0,7%	**76,9%**	0,0%	21,9%	100,0%
	Section 3	0,1%	10,6%	0,8%	**70,7%**	17,9%	100,0%
	Section 4	0,0%	**73,0%**	0,1%	0,8%	26,1%	100,0%

Table 7 gives the average number of references in a sentence calculated for each section type. This table clearly shows the difference in the average number of references between the different sections. The Introduction contains a very high number of references (more than one reference per sentence) and the same remains true for all the journals we examined. We see some slight differences between the journals: PLOS Computational Biology has less than one reference per sentence in the Introduction and 0.39 in the Discussion while all the other journals have more than 1.1 and 0.46 respectively. We note also that PLOS Medicine has the highest values for each section. However, these differences between the journals are not significant.

Table 7: Average number of references per sentence, by IMRaD section

Journal	Introduction	Methods	Results	Discussion
PLOS Biology	1,16	0,21	0,28	0,50
PLOS Comp. Biology	0,97	0,24	0,23	0,39
PLOS Genetics	1,14	0,23	0,29	0,47
PLOS Medicine	1,23	0,27	0,31	0,57
PLOS Negl. Trop. Diseases	1,11	0,22	0,18	0,51
PLOS Pathogens	1,21	0,20	0,25	0,55
PLOS ONE	1,13	0,19	0,21	0,53
All PLOS Journals	1,13	0,20	0,23	0,52

Distribution of References at the Sentence Level

Figure 1 presents the normalized distributions of references throughout the texts for the PLOS journals. The horizontal axis presents the text progression from 0 to 100 percent based on the segmentation into sentences. The vertical axis gives the average percentage of the total number of references at a given point of the text for each corpus. The vertical lines in the graph indicate the average positions of the section boundaries calculated separately for each corpus. Articles following the IMRaD structure with three sections (two sections combined in one) were not taken into consideration for the section boundaries.

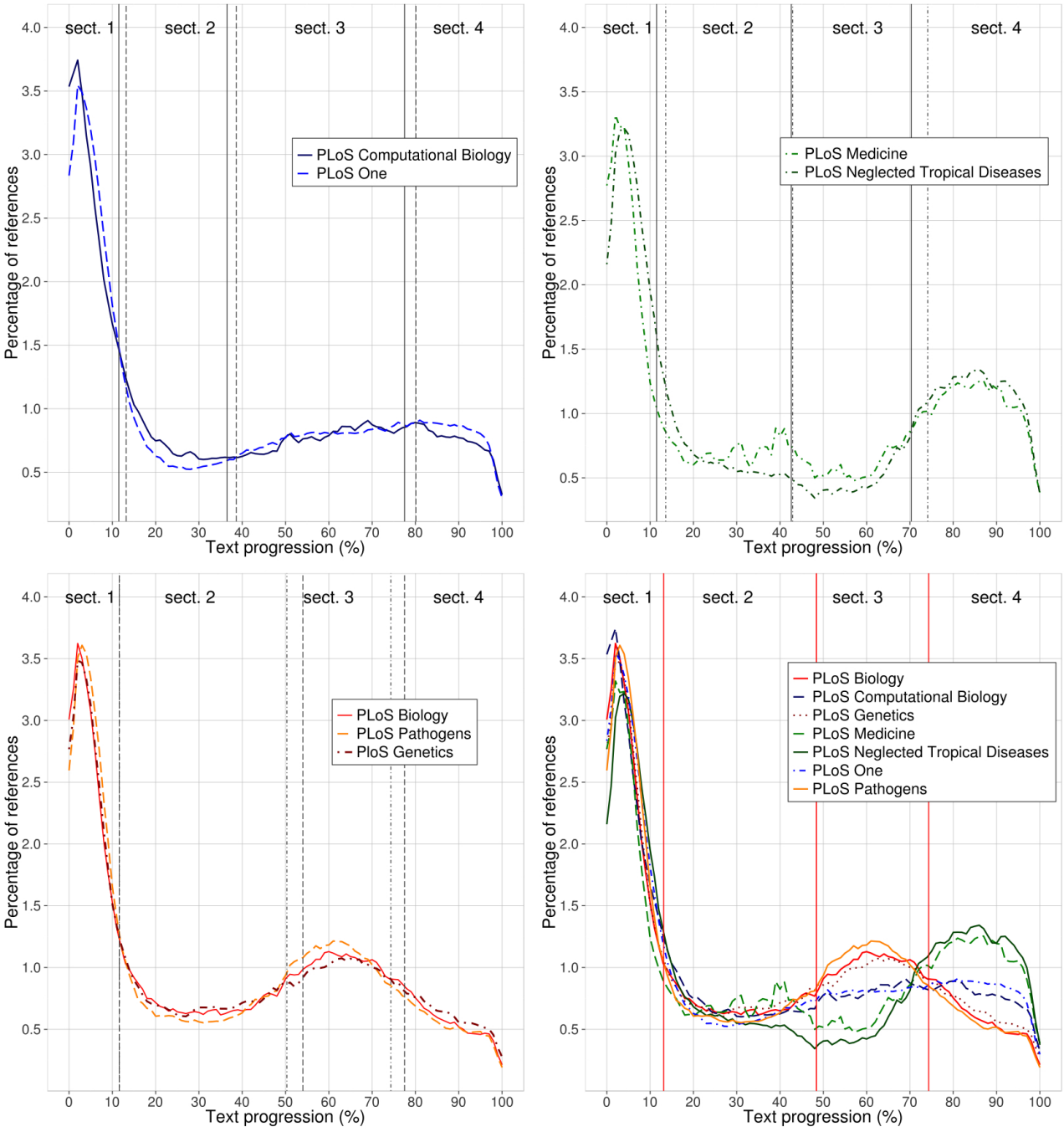

Figure 1: Distribution of references according to the text progression

We can observe that the first 10 percent of the texts in these corpuses contain relatively large number of references. These results are consistent with what might be expected: references are more concentrated in the first section which is the introduction in a vast majority of cases.

The curves we observe fall into 3 different types. The comparison of Table 6 with the curves on Figure 1 shows that the distributions of references are similar in the sets of journals having the same structure of section titles. In fact, Figure 1 shows that section 2, which according to Table 6 corresponds to the Method in a majority of articles in *PLOS Medicine* and *PLOS Neglected Tropical Diseases*, contains less references than the other sections. On the other hand, Table 6 shows that the Method section tends to be at the end of the articles for *PLOS Biology, PLOS Genetics and PLOS Pathogens*. This is consistent with the distribution of references on Figure 1 where we can observe that the fourth section contains a smaller proportion of

references than the first three sections. These observations suggest that if we take into account the variations in the positions of sections the distribution of references could be very stable and nearly invariant.

Distribution of References for the reordered IMRaD structure

To study the distribution of references independently of the order in which the sections of the IMRaD structure appear in the texts, we have reordered the sections in all articles to obtain a unique sequence: Introduction, Method, Result, Discussion. The articles, in which the sections were reordered, were then used to produce the new distribution of references. Figure 2 shows the results for the seven PLOS journals. The Introduction sections contain a relatively large number of references, with a bigger concentration in the first part of the Introduction. The Method section is characterized by a relatively smaller number of references which grows bigger towards the end of the Results section and gets higher again the Discussion section.

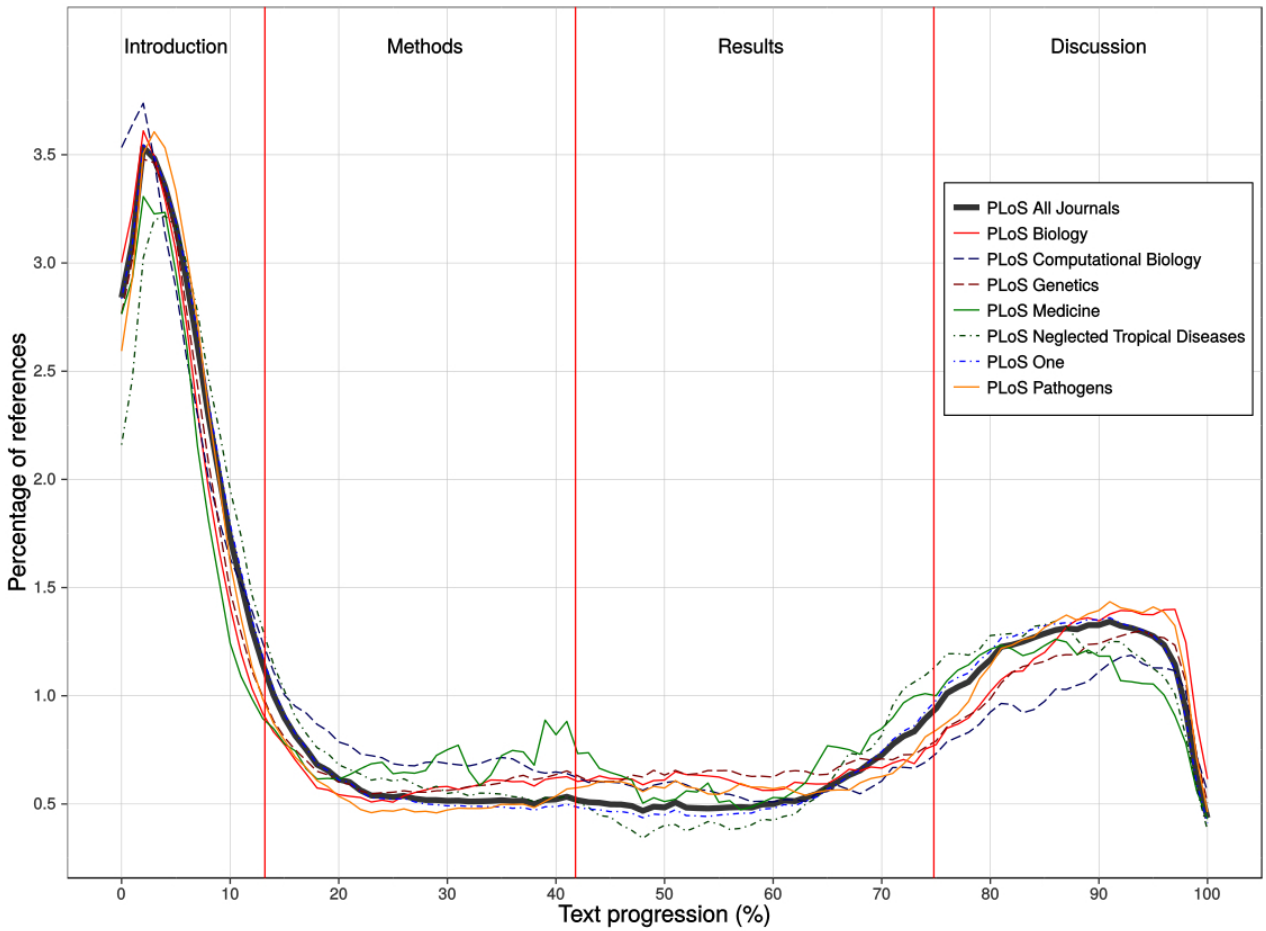

Figure 2: Distribution of references along the IMRaD structure

Table 8 provides the Pearson's correlation coefficients for all journal pairs. The correlations are significant at the level of 0.001 ($p<0.001$). There is a strong positive correlation ($r>0.9$) between the journal pairs, as well as between each journal and the values for the entire corpus ("PLOS All journals"). The smallest correlation is 0.908 between PLOS Computational Biology and PLOS Neglected Tropical Diseases.

Table 8: Pearson's correlation coefficients of the number of references at each point of the text progression from 0 to 100 (N=101).

	PLOS Biology	PLOS Comp. Biology	PLOS Genetics	PLOS Medicine	PLOS N. T. Diseases	PLOS Pathogens	PLOS ONE	PLOS All Journals
PLOS Biology	1	0,967	0,993	0,962	0,927	0,986	0,976	0,982
PLOS Comp. Biology	0,967	1	0,967	0,943	0,908	0,944	0,956	0,962
PLOS Genetics	0,993	0,967	1	0,966	0,946	0,991	0,984	0,989
PLOS Medicine	0,962	0,943	0,966	1	0,935	0,957	0,964	0,967
PLOS Negl. Trop. Diseases	0,927	0,908	0,946	0,935	1	0,961	0,982	0,977
PLOS Pathogens	0,986	0,944	0,991	0,957	0,961	1	0,989	0,991
PLOS ONE	0,976	0,956	0,984	0,964	0,982	0,989	1	0,999
All PLOS journals	0,982	0,962	0,989	0,967	0,977	0,991	0,999	1

Our results suggest that the distribution that we obtain is discipline-independent, as there exists a very strong correlation ($r>0.96$) between the general journal PLOS ONE and the other six journals that are domain-specific. On the whole, the high values of the correlations strongly suggest invariance in the distribution of references along the IMRaD structure.

Age Distribution of References

The distribution of references can also be considered with respect to the publication years of cited documents. As the IMRaD structure fixes the framework for the argumentative pattern of the text, we can expect that it also influences the age distribution of the cited documents along the sections.

The age of each reference is calculated as the difference between the year of the publication of the article containing the reference and the year of the cited document. Table 9 provides the average values for each section of the papers. We observe that:

- The discussion section tends to contain more recent references compared to the other sections.
- The oldest references can be observed in the Method section, followed by the Introduction.
- The age of references can vary considerably between the journals, for example PLOS Medicine tends to cite rather recent sources while PLOS Neglected Tropical Diseases cites, on average, older references. The difference between the two journals is very important in the Introduction and Method sections where the average age differs by about 3 years.
- For PLOS Genetics and PLOS Biology the differences between the four sections are rather small: all the values are between 6.24 and 7.58 for PLOS Genetics and between 6.82 and 7.93 for PLOS Biology, while for all the other journals the intervals are larger.

Table 9: Average age of references in the IMRaD structure

Journal	Introduction	Methods	Results	Discussion
PLOS Biology	7.83	7.93	7.09	6.82
PLOS Comp. Biology	7.80	8.96	7.40	7.29
PLOS Genetics	7.23	7.58	6.71	6.24
PLOS Medicine	6.57	7.71	8.18	6.34
PLOS Negl. Trop. Diseases	9.42	10.11	8.74	8.67
PLOS Pathogens	7.41	8.25	6.93	6.53
PLOS ONE	8.40	9.21	7.95	7.61
All PLOS journals	*8.22*	*8.97*	*7.66*	*7.44*

Figure 3 shows that in the first group of journals, PLOS Computational Biology and PLOS One, the oldest references are found in the beginning of the introduction and around the middle of the Method section. Then the age of the references steadily diminishes along the Result and Discussion sections. The same observations are valid for the second group of journals, PLOS Biology, PLOS Pathogens and PLOS Genetics. However, in these journals the differences between the sections are smaller as we could already observe in Table 9. The third group of journals, PLOS Medicine and PLOS Neglected Tropical Diseases, differ strongly from the other journals namely by the Introduction section where the references are more recent than the ones in the Method and Result sections. Here again, the age of references diminishes progressively along the last two sections.

The fourth graph on Figure 3 shows the distribution of age for each section for PLOS journals taken altogether. The beginning of the Introduction, in which the references have the highest density, contains also, on average, the oldest references of the article. The Method and the Result sections contain the smallest number of references, and the age of references diminishes progressively from the middle of the Method section to the end of the article. The Discussion section generally contains the most recent references.

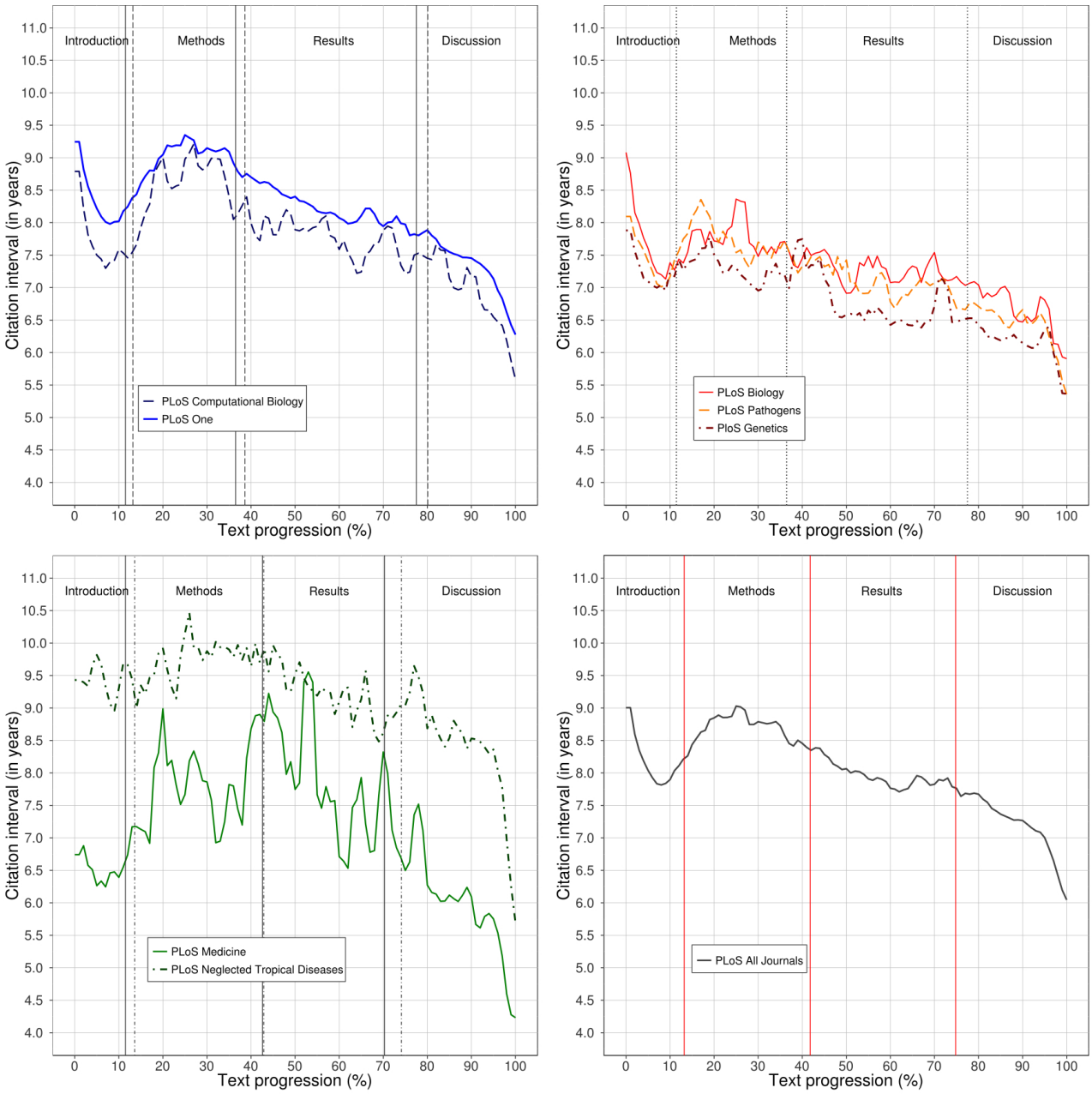

Figure 3: Age of references in the IMRaD structure

Table 10 presents the Pearson's correlation coefficients between all journal pairs for the age of references. The correlations are significant at the 0.01 level ($p<0.01$). The highest correlations exist between PLOS ONE and PLOS Pathogens ($r=0.853$) and between PLOS ONE and PLOS Computational Biology ($r=0.891$). Several journal pairs present relatively small correlations ($r<0.6$).

Table 10: Pearson's correlation coefficients for the average years at each point of the text progression from 0 to 100 (N=101)

	PLOS Biology	PLOS Comp. Biology	PLOS Genetics	PLOS Medicine	PLOS N. T. Diseases	PLOS Pathogens	PLOS ONE
PLOS Biology	1	0,707	0,700	0,394	0,558	0,751	0,790
PLOS Comp. Biology	0,707	1	0,653	0,554	0,652	0,721	0,891
PLOS Genetics	0,700	0,653	1	0,379	0,595	0,779	0,760
PLOS Medicine	0,394	0,554	0,379	1	0,479	0,456	0,562
PLOS N. T. Diseases	0,558	0,652	0,595	0,479	1	0,620	0,729
PLOS Pathogens	0,751	0,721	0,779	0,456	0,620	1	0,853
PLOS ONE	0,790	0,891	0,760	0,562	0,729	0,853	1

Discussion

Our hypothesis was that the structure of articles, represented by the IMRaD sections, affects the distribution of references in scientific articles. This cognitive hypothesis invites us to consider the relationship between the argumentative purpose of each section and the distribution of references. Our results demonstrate a strong relationship between the density of references and the argumentative structure of papers. These results show general trends and overcome some of the limitations of previous studies conducted manually and on much smaller datasets. We have processed the entire PLOS corpus up to September/October 2012 – for a total of about 45,000 articles – thus eliminating possible bias related to the choice of particular articles in the dataset to construct a representative sample as well as to a single scientific domain.

The distributions of references show some intrinsic properties of the IMRaD sequence and the structural differences between the sections. Moreover, we can now characterize each section by two types of distributions of the references: position along the text and age. One application of these results can be the automatic categorization of sections, irrespective of section titles.

The observed densities of references in the different sections of the IMRaD structure confirm the results reported by previous studies, which have shown that the first section, and more specifically the first 15% of the text, contains the highest density of references (Cano, 1989). In addition to that, our results clearly provide evidence that the distribution along the first 15% of the text is not homogeneous. High-density peaks are situated around 4% of the text progression, after which the density of references diminishes monotonously in the first section (Introduction). Previous studies on the IMRaD structure (Voos & Dagaev, 1976) and on the sections of scientific papers in general (Hu et al., 2013) report values for the average densities per section that are consistent with the overall distribution along the IMRaD structure (Figure 3). However, to our knowledge, this is the first study that shows the distribution of references along the entire text, thus accounting for the internal structure of sections.

The results on the age of references along the IMRaD sequence can be explained in light of the purpose of each section. The introductory section most often contains a literature review or a historical perspective of the research problem, from older to more recent sources. The Methods and Results sections refer to works that are already established and technologically mature as methodologies and reference data. This seems to be the reason for the relatively older references in these two sections. As McCain and Turner (1989) noted, "In the case of methods and materials innovations, even after the researcher is notified of their availability, it may take time to develop the skills to use the new technologies." Finally, the Discussion

section, that contains the more recent references, is the place where new perspectives and further research are given, as well as comparisons with other recent works in the same field.

One of the specificities of our approach is the fact that we work at the sentence level of the text. Previous studies in this field have measured the position of references in terms of number of words (Hu et al., 2013) or string lengths (Cano, 1989). From a linguistic point of view, sentences are relatively independent textual units that are natural building blocks of paragraphs. This makes them suitable as basic units to model text progression. The sentence-level distance implies:

- independence of word order, i.e. independence of the localization of a reference within a sentence;
- independence of sentence length and particular phrasing of ideas within a sentence.

Since all the seven journals that we analyzed in this study are published by PLOS, the question could be raised whether studying journals by other publishers would give similar results. For example, other more heterogeneous datasets can be analyzed such as PubMed, arXiv, or iSearch (Lykke et al., 2010). We note, however, that the PLOS corpus is multidisciplinary in nature, covering a wide variety of subjects in all natural, medical, and social sciences. Also, this study could be completed by other analyses per discipline, in order to investigate how specific disciplines differ from the general trend that we identified in Figure 2.

As a characteristic of the dataset, the values obtained for the number of in-text references (77.76) and items in the bibliography of research articles (48.94) -- see Table 3--, are consistent with the study of Ding et al. (2013) who also reports a ratio of about 1.6 between references and bibliography items on a corpus of 866 articles from the *Journal of the American Society for Information Science and Technology*. In their study, the average number of bibliography items is slightly lower: 37.52 per document. This might be due to differences in the content of journals or to the relatively small size of the sample.

We also observed that almost all of the articles in the corpus contain the four main sections of the IMRaD sequence. This may be related to editorial practices in the PLOS journals and therefore may not be true for other datasets. However, our aim is to characterize the IMRaD structure, which is predominant in scientific writing. For this we have limited our study to such papers, eliminating a small number of papers (about 1%) having other structures. Other rhetorical structures that exist remain to be analyzed. Our major conclusion is thus that given the fixed order of the IMRaD structure (corrected for inverted sections), the distribution of references along the text seems quite stable if not in fact invariant.

Conclusion

In this paper, we have measured the distribution of references along the text of scientific articles using sentences as the counting unit. We also showed that this distribution is quite stable and maybe even invariant if we take into account the changes that occur in some journals in the positions of the different sections in the text of the articles. By taking into account the various sections of the documents and reordering them into the IMRaD structure where necessary, it has been possible to link the references with their position in the text and better characterize the kinds of references according to the function of the section in which they appear. For example, it is plausible that the function of references found in the introductory section differs from that of references mentioned in the Method section or in the conclusion.

By processing the text at the sentence level and studying the order of the sections in the papers, we were able show that the distribution of references is closely related to the argumentative structure of papers. Furthermore, we provided evidence that this distribution is shared by papers across different domains, as it

remains the same for both domain-specific journals, such as PLOS Biology, and a general journal PLOS ONE that covers eleven subject areas in different disciplines, among which Biology and Life Sciences, Engineering and technology, Physical sciences and Social sciences.

We have characterized references along two criteria: position in sections and age of sources. As the four section types of the IMRaD sequence give the rhetorical framework of articles, the distribution of references along these types are important factors to consider when studying citation functions and citation acts. In this way, the results on the distributions provide insights into the different roles citations have in the scholarly communication process. The rhetoric of articles, represented by the IMRaD structure, affects the distribution of references in scientific articles. This cognitive hypothesis invites us to consider the relationship between the argumentative purpose of each section and the distribution of references. Similarly, we showed that the age of references vary considerably between the different PLOS journals. However, for each of these journals, the Method section tends to contain the oldest references while it is the section containing the smallest number of references. This suggests, again, that references found in different sections have a different purpose.

Using this method to localize references along the text, we plan in future work to focus on automatic citation context analysis using linguistic markers, and to examine other possible correlations that might exist between the position in the text and the nature of the references, their publication year or the subject category of the reference journals. The construction of a corpus of citation contexts with links to positions in the IMRaD structure can be envisaged through sentence extraction. Such data can be exploited in order to study citation contexts to propose a comprehensive typology of citation acts. For example, from the perspective of citation categorization (a study of lexical distributions in citation contexts was proposed by Bertin and Atanassova 2014). Such studies are also of interest in the field of Information Retrieval (Mayr, 2008 & 2013) and for the study of citation contexts. Our approach could contribute to the analysis of citation behavior (Cano 1989) as well as to the categorisation of the nature and functions of citation (Cronin 1981; 1982; 1984; 2004). The invariance of the distribution of references could be studied for other corpus and fields. Other phenomena that are related to this invariance have to be examined, such as the nature of citations and more specifically the acts of citation. We could, for example, find that papers cited in more than one section have different characteristics than those present in only one section. We thus think that a better understanding of the dynamics of citations can be obtained by analysing in more details the position of references along the papers at the sentence level.

Acknowledgments

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

References

Agarwal, S. and Yu, H. (2009). Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion, Bioinformatics 25(23): 3174-3180.

ANSI (2012). JATS: Journal Article Tag Suite. ANSI/NISO Z39.96-2012, 9 August 2012. National Information Standards Organization (NISO).

Barron, J. (2006). The Uniform Requirements for Manuscripts Submitted to Biomedical Journals Recommended by the International Committee of Medical Journal Editors. Chest, 129(4): 1098–1099.

Bertin, M. & Atanassova, I. (2012). Semantic Enrichment of Scientific Publications and Metadata: Citation Analysis Through Contextual and Cognitive Analysis. D-Lib Magazine, 18

Bertin, M., Atanassova, I., Lariviere, V., & Gingras, Y. (2013). The Distribution of References in Scientific Papers: an Analysis of the IMRaD Structure. In proceeding of: 14th International Society of Scientometrics and Informetrics Conference, 1 : 591-603

Bertin, M. & Atanassova, I. (2014). A Study of Lexical Distribution in Citation Contexts through the IMRaD Standard Bibliometric-enhanced Information Retrieval Workshop at European Conference on Information Retrieval, 1143, 5-12

Cano, V. (1989). Citation behavior: Classification, utility, and location, Journal of the American Society for Information Science 40(4): 284-290.

Carter, R.; Funk, K. and Mooney, R. (2012). The Front Matters: Capturing Journal Front Matter with JATS. Journal Article Tag Suite Conference (JATS-Con) Proceedings 2012. Bethesda (MD): National Center for Biotechnology Information (US); 2012.

Cronin, B. (1981). Agreement and Divergence on Referencing Practice, Journal of Information Science 3(1): 27-33.

Cronin, B. (1982). Norms and functions in citation: The view of journal editors and referees in psychology, Social Science Information Studies 2(2): 65-77.

Cronin, B. (1984). The citation process. The role and significance of citations in scientific communication, London: Taylor Graham, 1984 1.

Cronin, B. (2004). Normative shaping of scientific practice: The magic of Merton, Scientometrics 60(1): 41-46.

Day, R. A. and Gastel, B. (2006). How to write and publish a scientific paper. Greenwood Press, Westport, CT.

Ding, Y.; Liu, X.; Guo, C. and Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis, Journal of Informetrics 7(3): 583-592.

Hu, Z.; Chen, C. and Liu, Z. (2013). Where are citations located in the body of scientific articles? A study of the distributions of citation locations, Journal of Informetrics 7: 887 - 896.

Kucer, S. L. (1985). The Making of Meaning Reading and Writing as Parallel Processes, Written Communication 2 : 317-336.

Liu, S. and Chen, C. (2013). The differences between latent topics in abstracts and citation contexts of citing papers. Journal of the American Society for Information Science and Technology, 64(3), 627-639.

Lykke, M., Larsen, B., Lund, H., Ingwersen, P. (2010). Developing a Test Collection for the Evaluation of Integrated Search. In Gurrin, C. et al. eds., Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010, Proceedings. Berlin, Springer p. 627-630. (Lecture Notes in Computer Science; 5993).

Maričić, S.; Spaventi, J.; Pavičić, L. & Pifat-Mrzljak, G. (1998). Citation context versus the frequency counts of citation histories, Journal of the American Society for Information Science, Wiley Online Library, 49, 530-540.

Mayr, P.; Mutschke, P. & Petras, V. (2008). Reducing semantic complexity in distributed digital libraries: Treatment of term vagueness and document re-ranking Library Review, Emerald Group Publishing Limited, 57, 213-224.

Mayr, P. (2013). Relevance Distributions Across Bradford Zones: Can Bradfordizing Improve Search ? 14th International Society of Scientometrics and Informetrics Conference, 2, 1493-1505

McCain, K. W. & Turner, K. (1989). Citation context analysis and aging patterns of journal articles in molecular genetics Scientometrics. Akadémiai Kiadó, co-published with Springer Science+ Business Media BV, Formerly Kluwer Academic Publishers BV, 17, 127-163.

Meadows, A. (1985). The scientific paper as an archaeological artefact, Journal of information science 11(1): 27-30.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. Social studies of science, 5(1), 86-92.

Mourad, G. (2001). SegATex et CitaRE, Computational Analysis of typographic signs for the segmentation of text and Automatic Extraction of Citations (in French), University of Paris-Sorbonne.

Nakayama, T.; Hirai, N.; Yamazaki, S. and Naito, M. (2005) Adoption of structured abstracts by general medical journals and format for a structured abstract, Journal of the Medical Library Association, 93(2), 237-242.

Oriokot, L.; Buwembo, W.; Munabi, I. and Kijjambu, S. (2011). 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.

Small, H.(2011). Interpreting maps of science using citation context sentiments: a preliminary investigation. Scientometrics, 87(2): 373-388.

Sollaci, L. B. and Pereira, M. G. (2004). The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey, Journal of the Medical Library Association 92(3): 364.

Teufel, S.; Siddharthan, A. and Tidhar, D. (2006). Automatic classification of citation function, COLING/ACL 2006 - EMNLP 2006: 2006 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference: 103-110.

Voos, H., & Dagaev, K. S. (1976). Are All Citations Equal? Or, Did We Op. Cit. Your Idem?. Journal of Academic Librarianship, 1(6), 19-21.

Wan, X. & Liu, F. (2014) Are all literature citations equally important? Automatic citation strength estimation and its applications. Journal of the Association for Information Science and Technology, Wiley Online Library.