# The linguistic patterns and rhetorical structure of citation context: an approach using n-grams

Marc Bertin[1], Iana Atanassova[2], Cassidy R. Sugimoto[3] andVincent Lariviere[4]

[1] *bertin.marc@gmail.com*
Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Québec à Montréal, CP 8888, Succ. Centre-Ville, Montreal, QC. H3C 3P8 (Canada)

[2] *iana.atanassova@univ-fcomte.fr*
Centre de Recherche en Linguistique et Traitement Automatique des Langues "Lucien Tesnière", Université de Bourgogne Franche-Comté, Besançon 25000 (France)

[3] *sugimoto@indiana.edu*
School of Informatics and Computing, Indiana University Bloomington, IN 47405 (USA)

[4] *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, Montreal, QC. H3C 3P8 (Canada)

## Abstract

Using the full-text corpus of more than 75,000 research articles published by seven PLOS journals, this paper proposes a natural language processing approach for identifying the function of citations. Citation contexts are assigned based on the frequency of n-gram co-occurrences located near the citations. Results show that the most frequent linguistic patterns found in the citation contexts of papers vary according to their location in the IMRaD structure of scientific articles. The presence of negative citations is also dependent on this structure. This methodology offers new perspectives to locate these discursive forms according to the rhetorical structure of scientific articles, and will lead to a better understanding of the use of citations in scientific articles.

## Introduction

An important issue for the sociology of science and knowledge is to empirically understand the relationship between past and current scholarship (Hargens, 2000). This is most often done by means of citation analysis, which traces links of interaction from one piece of scholarship to another and, over time, provides an empirical foundation for the process of knowledge accumulation. As Derek de Solla Price noted, "the most obvious manifestation of this scholarly bricklaying is the citation of references" (1963, p. 64-65). However, databases such as the Science Citation Index—created more than 50 years ago—provide only a binary distinction of the link between two articles. That is, they demonstrate whether a work was cited in another, but not where in the article, nor in what way. This is not without some advocacy for improvement: Lipetz, for example, argued in the mid-1960s that citation indices should not only indicate that something was cited, but also include the "disposition of the scientific contribution of the cited paper in the citing paper" (Lipetz, 1965). This was not adopted by the Science Citation Index or in any other major citation index for the next few decades. Therefore, early citation context studies were reliant upon manually annotated texts.

Early sociological studies sought to provide an empirical basis for understanding the norms and social functions of citations (e.g., Kaplan, 1965; Gilbert, 1977). For example, Garfield (1964) provided one of the first enumerations of reasons for citing. The list of fifteen reasons included "paying homage to pioneers", criticizing previous work, authenticating data and classes of fact, and disputing priority claims. Many scholars followed suit, seeking to create

classifications, typologies, and schema that encompass both the function of and motivation for citations (e.g. Moravcsik & Murugesan, 1975; Swales, 1986; White, 2004; Teufel, Siddharthan, & Tidhar, 2006; Teufel, Siddharthan, & Tidhar, 2009; Jörg, 2008; Chubin & Moitra, 1975; Garfield, 1964; Garfield et al. 1972; Small, 1982; Cronin, 1984; Liu, 1993; Small, 1982). Given that most of these were constructed manually, the analyses on which they were built were often fairly small in scale, tracing, for example, the citation context of a single paper (e.g., Anderson & Sun, Sieweke, 2014) or journal (Spiegel-Rosing, 1977; Halevi & Moed, 2013).

With the advent of the digital era, large-scale datasets containing the full-text of scholarly documents have become available and initiatives have been undertaken to provide appropriate mark-up for citation context analysis (e.g., Fujiwara & Yamamoto, 2015; Giles et al., 1998; Peroni & Shotton, 2012). Hence, over the last few year, we are witnessing an increasing body of literature on the citation context of scientific papers (e.g. Boyack, Small, & Klavans, 2013; Zhao & Strotmann, 2014; Bertin, Atanassova, Larivière, & Gingras, 2013; Bertin & Atanassova, 2014; Catalini, Lacetera & Oettl, 2015). Due to their extensive XML mark-up, the corpi from PLOS (e.g., Bertin, Atanssova, Gingras, & Lariviere, 2016), PubMed Central (Liu et al., 2014; Elkiss et al., 2008; Callahan, Hockema, & Eysenbach, 2010) and Elsevier (Boyack, Small, & Klavans, 2012), have been particularly useful for these analyses.

Citation context has been operationalized in several ways. For many scholars, context implies the position within the text (e.g., Bertin, Atanssova, Gingras, & Lariviere, 2016) or in relation to other references (Elkiss et al., 2008; Liu & Chen, 2012; Gipp & Beel, 2009; Boyack et al., 2012; Callahan, Hockema, & Eysenbach, 2010). Furthermore, scholars have sought to understand the function of the references by the location in text alone, or in relation to frequency (e.g., Ding, Liu, Guo, & Cronin, 2013; Marici et al., 1998). Citation context has also been done to relate not to position or relation to other references, but the semantics surrounding the reference (e.g., Siddarthan & Teufel, 2007).

Many schemas for categorizing citations have been proposed, but most lack a clear operationalization (e.g., distinctive linguistic markers) that would allow for large-scale automated analyses (Sula & Miller, 2014). That is, categorizations fail to provide ways to operationalize at scale, distinguishing, for example, between "paying homage to pioneers" from "providing background reading" (Garfield, 1964/1970). Scholars have sought to bridge the gap between manual and automatic classification, by use of validation studies demonstrating high convergence between citations manually classified with those automatically classified (e.g., Liu et al., 2014; Teufel, Siddharthan, & Tidhar, 2006). In the case of Teufel, Siddharthan and Tidhar (2006), the manual annotation was used in order to generate and confirm cue words: that is, meta-discourse that would allow for automatic classification of the citation context. These words are often taken from the text surrounding the citation, which has been termed citation summaries (Elkiss et al., 2008), citances (Nakov, Schwartz, & Hearst, 2004), citing statements (O'Connor, 1982) or—more broadly—the citation context (Small, 1979).

Many different units have been parsed: for example, the sentences preceding and following the citation (O'Connor, 1982), noun phrases in up to three sentences that surround the citation (Schneider, 2006), and n-grams (Sula & Miller, 2014). Verbs found in the citation context are particularly useful in providing insight on the relationship between citing and cited documents, as they are related to the rhetorical context (Hopper, 1987; Sakita, 2002; Bloch, 2010). Sentiment analysis can also be conducted, to gauge whether the reception of the cited

article is positive, negative, or neutral (Teufel et al., 2006). Using a combined approach, Sula and Miller (2014) utilized both n-grams and sentiment analysis to place citation in context along a spectrum from negative to positive (which they referred to as polarity) as well as identifying the location of the reference within the text. A similar approach will be taken in this analysis, to identify, at scale and across disciplines, n-grams and sentiment.

Done at scale, citation context analysis has potential utility for summarization (Nanba & Okumura, 1999; Mohammad et al., 2009), information retrieval (Liu et al., 2014; Bradshaw, 2003; O'Connor, 1982), clustering (Boyack, Small, & Klavans, 2012), disambiguation, and entity recognition (Nakov, Schwartz, & Hearst, 2004). It also allows scholars to return to the original promise of citation context analysis: that is, to uncover the core elements, or concept symbols, represented in the cumulative citation contexts of an item (Schneider, 2006; Mei & Zhai, 2008). This work focuses on the full text processing of paper and the linguistic phenomena around references by examining n-grams extracted from citation contexts. In order to contribute to this newly revived discussion, this paper proposes a natural language processing approach for identifying the function of citations. We assign labels to in-text citations—i.e., citations and their context, as they appear in the byline of scientific papers—based on the frequency of n-gram co-occurrences located near the citations. More specifically, our aim is to uncover the most frequent linguistic patterns—based on n-grams—found in the citation contexts of papers, and to assess whether these patterns vary according to their location in the rhetorical structure of scientific articles—defined here as the IMRaD structure (introduction, methods, results, discussion/conclusion)—using more than 75,000 research articles from PLOS.

## Methods

*Dataset*

The dataset contains all research articles appearing in the seven journals published by the Public Library of Science (PLOS). The research articles are available in XML format following the Journal Article Tag Suite (JATS) schema. We have processed the entire corpus from August 2003 up to September 2013, which contains 75,964 research articles, covering all fields of knowledge by primarily in the biomedical domain.

After parsing the documents, we identified all in-text citations and their corresponding bibliography items. This process involved the following two steps:

1. Identification of all *xref* elements of the XML documents that represent in-text citations;
2. Full-text processing of all sentences in order to identify any in-text citations that were missed by the first step using regular expressions. In fact, a few in-text citations in the corpus are present in the text but not identified as XML elements.

The research articles follow the IMRaD structure (i.e., Introduction, Methods, Results, and Discussion—with the literature and background incorporated into the Introduction), which is imposed by editorial requirements. About 97% of the articles in the corpus contain the four typical sections of the IMRaD structure. Each article was divided into these four sections according to the method described in Bertin, Atanassova, Larivière, and Gingras (2016) and each section was extracted an analyzed as a sub-dataset. Table 1 presents the number of research articles for each of the journals, the number of sentences containing citations, and the number of extracted 3-grams from these sentences.

**Table 1. Descriptive statistics on the PLOS dataset**

| Journal | Articles | Sentences containing citations | Extracted 3-grams |
|---|---|---|---|
| PLOS Medicine | 915 | 34,172 | 351,548 |
| PLOS Biology | 1,735 | 90,148 | 906,136 |
| PLOS Neglected Tropical Diseases | 1,867 | 72,970 | 709,595 |
| PLOS Computational Biology | 2,418 | 120,110 | 1,226,606 |
| PLOS Pathogens | 2,973 | 162,698 | 1,607,647 |
| PLOS Genetics | 3,402 | 184,878 | 1,820,127 |
| PLOS ONE | 62,654 | 2,474,710 | 23,934,434 |
| *Total* | *75,964* | *3,139,686* | *30,556,093* |

*N-grams computation and selection*

One approach to represent citation contexts is to use sequences of words, called n-grams (Cavnar et al. 1994), where *n* represents the number of words in sequence (typically $2<n<=5$). For this paper, we choose 3-grams, that is, those sequences of three words, and consider only 3-grams within sentence boundaries, as sentences are the natural building blocks of text and likely to include the context of a specific reference. Modelling text as a set of sentences as opposed to a sequence of words is better motivated from the linguistic point of view, as sentences are 1) textual units that can express meaning in a manner that is relatively independent from their context, and 2) they are used as basic units of text in a large number of works in applied linguistics (see e.g. Nenkova and MaKeown (2011) and Athar and Teufel (2012)).

However, if all 3-grams are considered, some of the information is duplicated, as shown in the example below:

**it was shown**
**was shown that**
**shown that the**

For this reason, we need further processing to reduce the size of the n-gram sets. To do this, we performed the POS-tagging of the sentences and selected only the n-grams of citation contexts that contain verbs. In general, verbs give important information about the nature of the relation between the article and the cited work. Polysemy is one possible problem when dealing with verbs, but in our case this phenomenon is reduced as we work specifically on citation contexts. By keeping only n-grams that contain verbs, we eliminate word patterns containing only nominal information like: "*In this paper", "the present article", "the result of"* etc. We note that, following this protocol, each occurrence of a verb in a sentence generates *n* different *n*-grams, except for cases where the verb is within *n* words from the beginning or the end of the sentence. Therefore, for each occurrence of a verb, we obtain between 1 and *n* n-grams. From this corpus, we extract a dataset which contains the sentences with in-text citations. For each sentence we obtain the set of 3-grams containing verbs, as well as its position in the article and in the section, the type of the section according to the IMRaD structure.

For each verb, we consider the set of 3-grams containing this verb, that we call a *class*.

## Results

As shown by Bertin and Atanassova (2014), the most common verbs in these data, by section, are *show, use, include, suggest* (Introduction), *use, perform, follow, obtain* (Methods), *use,*

*show, find, report* (Results), and *show, suggest, use, report* (Discussion). We also provide results for several verbs that carry specific meanings for the study of citation contexts: namely, *know, demonstrate, propose, calculate, describe, observe, agree,* and *disagree*. To examine sentiment, we also analyze positive and negative forms of the verbs. All results are presented according to the section(s) of the IMRaD structure in which they are most frequently found. The horizontal axis presents the text progression from 0% to 100% following the IMRaD structure. In the cases where the four sections of the IMRaD structure appear in different order—for example the methods presented after results—these have been reordered for coherence. The vertical axis provides the percentage of occurrences of each class relative to the total number of articles.

*Common verbs*

Figure 1 presents the verbs that that have a high frequency in the Introduction and Discussion sections—with higher frequencies in the former than in the latter—and a relatively small number of occurrences in the Methods and Results sections. The verbs most often found in this category includes classes such as <Show>, <Suggest>, <Find>, <Know>, <Demonstrate>, <Include> and <Propose>.

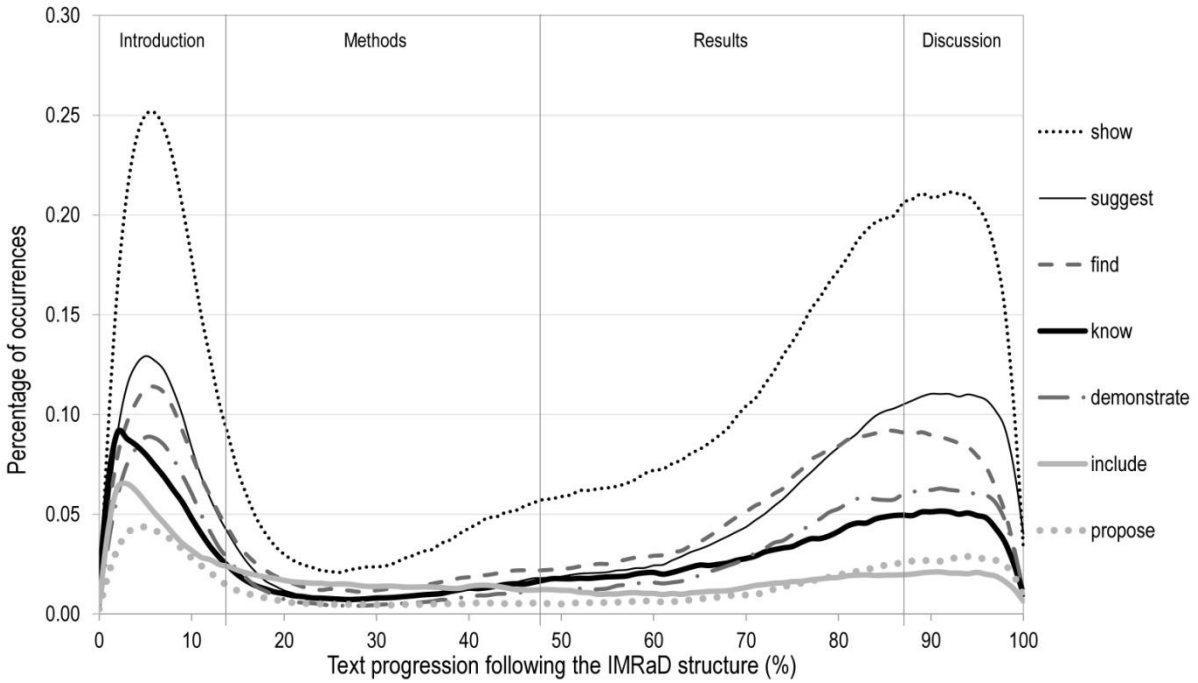


**Figure 1. Distribution of 3-grams following the IMRaD structure: Type 1-a**

Table 2 presents the verb classes and their cumulative percentages. We observe that for the classes <Show>, <Know>, and <Propose> the ten most frequent 3-grams account for more than 26% of all occurrences. For the other classes this number if lower, which means that they present more diversity in the 3-grams that compose them. For the class <Include> only 6.6% of the occurrences are covered by the ten most frequent 3-grams. This class contains a relatively very high number of 3-grams.

**Table 2. Cumulative percentages for the most frequent 3-grams: Type 1-a**

| **<Show>** | | **<Suggest>** | |
|---|---|---|---|
| been shown to | 6.52 | ,_, suggesting that | 4.03 |
| has been shown | 11.86 | has been suggested | 6.52 |
| shown to be | 14.91 | -RSB-_-RRB- ,_, suggesting | 8.78 |
| have shown that | 17.95 | been suggested that | 10.32 |
| have been shown | 20.06 | suggested that the | 11.69 |
| studies have shown | 21.68 | suggest that the | 12.80 |
| been shown that | 23.17 | have suggested that | 13.89 |
| shown that the | 24.28 | suggesting that the | 14.91 |
| was shown to | 25.36 | been suggested to | 15.87 |
| not shown -RRB-_-RRB- | 26.35 | suggests that the | 16.77 |
| **<Find>** | | **<Know>** | |
| found to be | 2.73 | is known to | 5.44 |
| found in the | 4.29 | are known to | 10.88 |
| was found to | 5.84 | known to be | 15.36 |
| we found that | 7.37 | is known that | 19.84 |
| been found to | 8.83 | also known as | 21.63 |
| ,_, we found | 10.27 | is well known | 23.32 |
| has been found | 11.59 | It is known | 24.55 |
| found that the | 12.74 | is known about | 25.74 |
| have been found | 13.86 | it is known | 26.75 |
| were found to | 14.84 | well known that | 27.63 |
| **<Include>** | | **<Demonstrate>** | |
| ,_, including the | *1.74* | has been demonstrated | 3.32 |
| -RSB-_-RRB- ,_, including | 3.47 | have demonstrated that | 6.64 |
| included in the | 4.23 | demonstrated that the | 9.42 |
| ,_, including a | 4.96 | studies have demonstrated | 11.48 |
| proteins ,_, including | 5.24 | been demonstrated to | 13.39 |
| factors ,_, including | 5.51 | been demonstrated that | 14.83 |
| ,_, which includes | 5.79 | been demonstrated in | 16.06 |
| ,_, including those | 6.06 | previously demonstrated that | 17.13 |
| were included in | 6.33 | have been demonstrated | 18.03 |
| genes ,_, including | 6.60 | -RSB-_-RRB- demonstrated that | 18.84 |
| **<Propose>** | | | |
| has been proposed | 6.41 | | |
| have been proposed | 12.81 | | |
| been proposed to | 16.19 | | |
| been proposed that | 19.55 | | |
| been proposed as | 22.21 | | |
| proposed to be | 23.81 | | |
| proposed that the | 25.34 | | |
| proposed as a | 26.78 | | |
| been proposed -LSB-_-LRB- | 27.88 | | |
| been proposed for | 28.76 | | |

The classes <Observe> and <Report> are also highly occurring in the Introduction and Discussion, except that their frequencies in the Discussion section are higher than in the Introduction section (type 1-b). Figure 2 presents the distributions for these two classes, and

Table 3 presents the most frequent 3-grams that belong to these classes and their cumulative percentages.

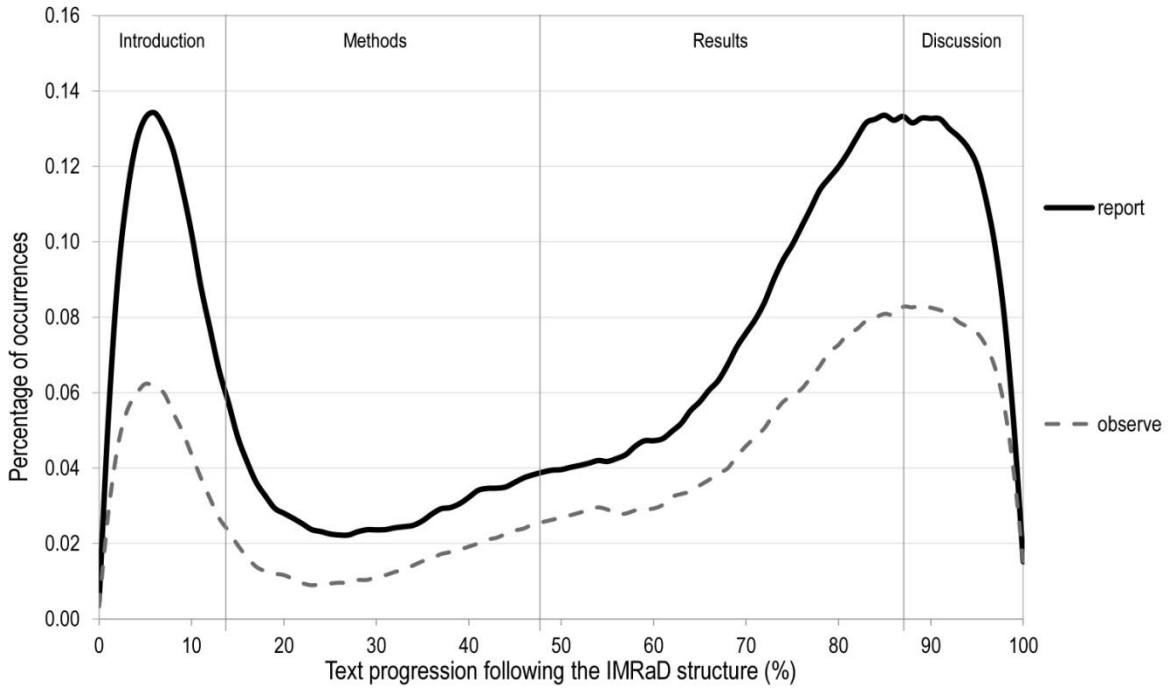


**Figure 2. Distribution of 3-grams following the IMRaD structure: Type 1-b**

**Table 3. Cumulative percentages for the most frequent 3-grams: Type 1-b**

| <Observe> | | <Report> | |
|---|---|---|---|
| observed in the | 2.05 | has been reported | 3.99 |
| has been observed | 4.11 | been reported to | 7.98 |
| been observed in | 6.07 | have been reported | 10.27 |
| have been observed | 8.03 | reported to be | 12.54 |
| was observed in | 9.09 | been reported in | 14.15 |
| ,_, we observed | 10.14 | been reported that | 15.55 |
| also observed in | 10.97 | previously reported -LSB-_-LRB- | 16.88 |
| also been observed | 11.54 | reported that the | 17.95 |
| were observed in | 12.07 | as previously reported | 18.87 |
| was also observed | 12.59 | been reported -LSB-_-LRB- | 19.62 |

The second main type of classes of n-grams contains classes that are characteristic of the Methods section (type 2). Figure 3 presents the distributions for the classes <Describe>, <Perform>, <Calculate> and <Obtain>. These classes have relatively high frequencies in the Methods section and low frequencies in the other sections. This means that these classes are used in a different manner than the types 1-a and 1-b, and they allow the expression of semantic relations specifically related to the Methods section in the rhetorical structure.

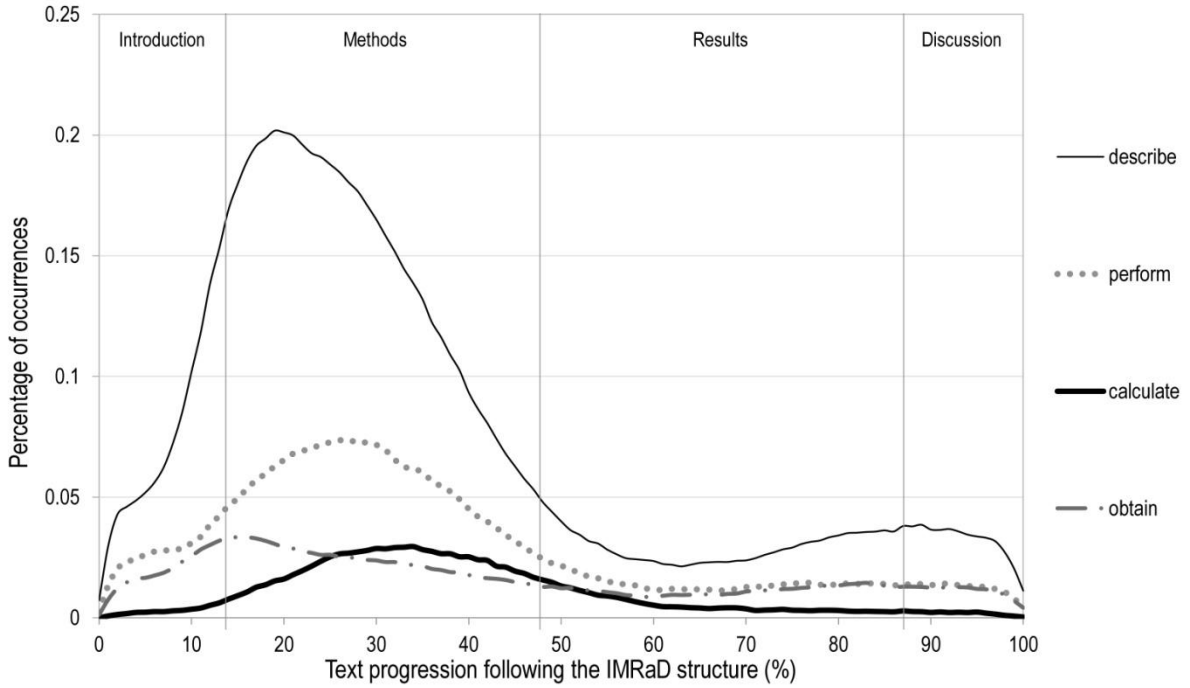


**Figure 3. Distribution of 3-grams following the IMRaD structure: Type 2**

Table 4 presents the most frequent 3-grams for these classes. The class <Describe> is represented by a relatively small number of 3-grams. As we can see in Table 4, ten of the 3-grams containing "describe" account for more than 40% of all occurrences of the verb in citation contexts.

**Table 4. Cumulative percentages for the most frequent 3-grams: Type 2**

| **<Describe>** | | **<Perform>** | |
|---|---|---|---|
| previously described -LSB-_-LRB- | 7.24 | was performed as | 4.54 |
| as previously described | 14.48 | performed as described | 8.87 |
| described previously -LSB-_-LRB- | 21.33 | performed as previously | 11.71 |
| as described previously | 26.93 | were performed as | 14.54 |
| as described -LSB-_-LRB- | 31.54 | was performed using | 17.02 |
| performed as described | 34.58 | were performed using | 18.53 |
| have been described | 36.31 | analysis was performed | 19.95 |
| as described in | 37.66 | performed using the | 21.35 |
| ,_, as described | 38.99 | ,_, we performed | 22.56 |
| has been described | 40.29 | assays were performed | 23.54 |
| **<Calculate>** | | **<Obtain>** | |
| calculated using the | 3.08 | were obtained from | 3.50 |
| were calculated using | 6.15 | obtained from the | 7.00 |
| was calculated using | 8.71 | was obtained from | 10.11 |
| to calculate the | 10.89 | were obtained by | 11.60 |
| was calculated as | 12.79 | was obtained by | 12.59 |
| used to calculate | 14.59 | to obtain a | 13.25 |
| -RRB-_-RRB- was calculated | 16.32 | -RRB-_-RRB- were obtained | 13.87 |
| ,_, we calculated | 17.66 | were obtained using | 14.47 |
| was calculated by | 18.92 | obtained from a | 15.06 |
| calculated from the | 20.14 | to obtain the | 15.64 |

The third main type of classes of n-grams contains classes that have relatively high frequencies in the Methods section, while also having high frequencies in the Introduction section (type 3). Such classes are for example <Use> and <Follow>. They contain expressions that appear especially in the Introduction and Methods sections and are relatively rare in the Results section. Figure 4 presents the distributions for these classes and Table 5 presents the most frequent 3-grams.

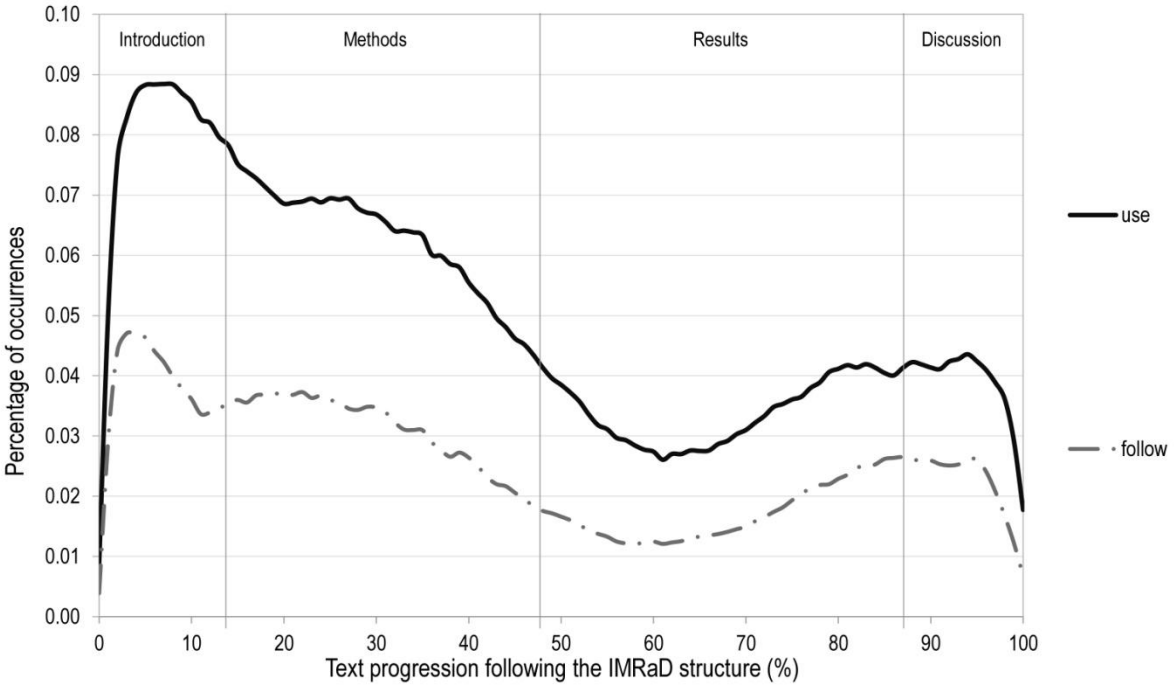


**Figure 4. Distribution of 3-grams following the IMRaD structure: Type 3**

**Table 5. Cumulative percentages for the most frequent 3-grams: Type 3**

| <Use> | | <Follow> | |
|---|---|---|---|
| was used to | 3.54 | ,_, followed by | 2.21 |
| ,_, we used | 6.45 | followed by a | 3.18 |
| We used the | 8.45 | as follows :_: | 4.11 |
| -RSB-_-RRB- was used | 10.26 | with the following | 4.80 |
| has been used | 11.72 | using the following | 5.32 |
| were used to | 13.16 | followed by the | 5.83 |
| been used to | 14.47 | ,_, following the | 6.28 |
| we used a | 15.69 | of the following | 6.66 |
| can be used | 16.86 | is followed by | 7.00 |
| be used to | 17.99 | -RSB-_-RRB- ,_, followed | 7.31 |

*Agreement and negation*

If we consider the distributions for the n-gram classes <Agree> and <Disagree>, we can observe the expression of agreement and disagreement between authors in the rhetorical structure. As shown in Figure 5, the class <Agree> is very frequent in the Discussion section and in general its frequency tends to grow steadily along the Methods and Results sections. This shows that agreement is expressed mostly at the end of a research article, especially in the Discussion section and towards the end of the Results section. Disagreement is less common in scientific discourse: the class <Disagree> is evenly distributed along the four sections of the IMRaD structure and has a very low frequency. Table 6 presents the top ten 3-

grams for the classes <Agree> and <Disagree> and their cumulative percentages. We can observe that the expression of agreement allows little variation in the linguistic means: ten of the 3-grams account for more than 71% of all occurrences of the verb "agree". In contrast, the class <Disagree> contains more variations and the top ten 3-grams account for only about 17% of all occurrences.

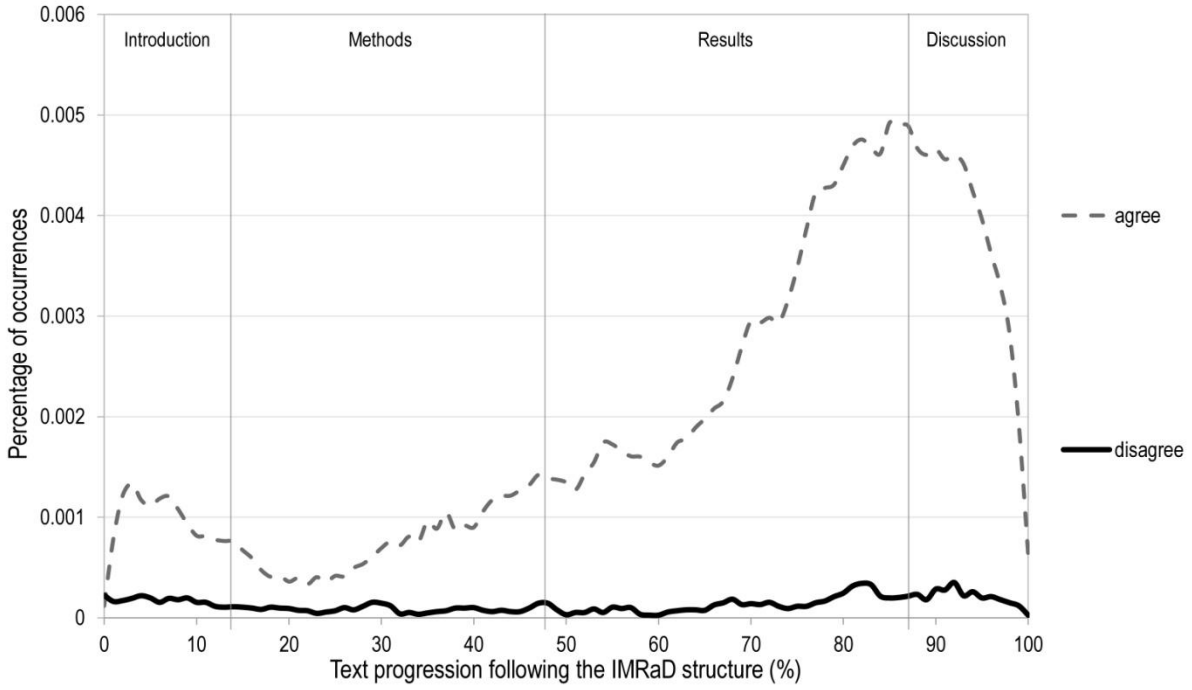


**Figure 5. Distribution of 3-grams following the IMRaD structure: authors' point of view**

**Table 6. Cumulative percentages for the most frequent 3-grams: <Agree> and <Disagree>**

| <Agree> | | <Disagree> | |
|---|---|---|---|
| is in agreement | 23.97 | is in disagreement | 3.21 |
| are in agreement | 47.95 | disagrees with the | 6.41 |
| results agree with | 60.98 | are in disagreement | 8.12 |
| This agrees with | 62.78 | there is disagreement | 9.72 |
| ,_, which agrees | 64.56 | disagree with the | 11.32 |
| was in agreement | 66.18 | results disagree with | 12.83 |
| which agrees with | 67.73 | to disagree with | 14.03 |
| were in agreement | 69.22 | disagree with a | 15.03 |
| These results agree | 70.38 | disagreed with the | 15.83 |
| findings agree with | 71.30 | disagree on the | 16.63 |

We have also examined the distribution of negations near citations. We extracted from the 30M 3-grams of the dataset all forms with negative word `not`. This simple example gives produce 196,926 forms but only 20,482 distinct discursive patterns. The first 30 percent of these forms concern only 38 3-grams and produce patterns like: *cannot be, did not find, is not clear, did not observe, did not show, did not affect, did not detect, was not detect*, etc. Figure 6 shows the distribution of n-grams that contain "not" along the IMRaD structure. We observe the highest frequency in the Discussion section, and the frequency grows steadily along the Results section. This distribution along the Results and Discussion sections in very

similar to that of the class <Agree> that we saw above. Additionally, there is a very high frequency in the Introduction section.

Figure 6 presents the distribution of different verbs with negative forms.

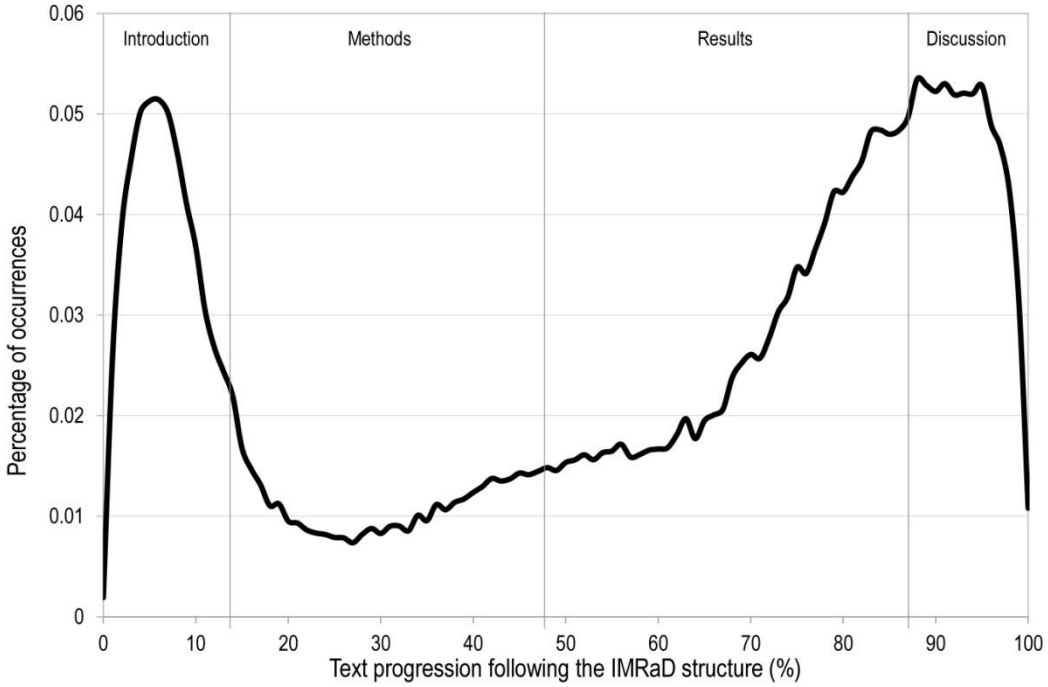


**Figure 6. Distribution of 3-grams following the IMRaD structure: the position of negations near citations**

Table 7 presents the top 20 3-grams containing negation. Apart from the verb "to be", negations are most frequently used with verbs like "show", "find", "know", "observe".

**Table 7. Cumulative percentages for the 20 most frequent 3-grams containing negations**

| | | | |
|---|---|---|---|
| data not shown | 3.23 | is not clear | 16.58 |
| has not been | 6.20 | did not observe | 17.17 |
| can not be | 8.70 | did not show | 17.75 |
| have not been | 10.37 | did not affect | 18.32 |
| may not be | 11.92 | do not have | 18.85 |
| could not be | 12.94 | has not yet | 19.37 |
| is not a | 13.76 | does not affect | 19.89 |
| did not find | 14.56 | is not the | 20.37 |
| is not surprising | 15.27 | is not required | 20.85 |
| is not known | 15.95 | does not appear | 21.32 |

## Discussion

Citations serve a central function in the scholarly communication system: representing both priority (Merton, 1957; 1961) and peer (Merton, 1998) recognition. Citations also function as a symbolic language of science, reflecting the underlying substance of and relationship among scientific documents (Small, 1978). This notion was codified in Small's (1978) theory of a concept symbol: that is, when cited documents become symbolic of the theories, concepts, or methods for which they are referenced. Despite the fundamental role of citations for science, a

single overarching model or theory of citation remains elusive (Cronin, 1981; Cronin, 1984; Small, 1982; Leydesdorff, 1998; Sugimoto, 2016). This is in no small part to the lack of robust citation context analyses. As Small noted, "a theory of citation must go beyond the binary presence or absence of a citation to the "comparison of the cited text with its context of citation in the citing texts" (Small, 2004, p. 76). Small began exploring the relationship between the context of a citation and the structure of scientific knowledge in the late 1970s (Small, 1979). More than a decade later, in a lecture in 1990, Thomas Kuhn implied that lexical analyses of citation contexts could yield insights into cognitive evolutions in scientific communities (Small, 2011). It is only with the access to full-text databases and computational power that we are able to begin to draw sophisticated analyses from the body of scientific literature.

Studies of location of references have demonstrated that references are not distributed equally across the text, but rather concentrate in the Introduction and, to a lesser extent, in the Discussion (Hargens, 2000; Voos & Dagaev, 1976; Bertin, Atanassova, Larivère & Gingras, 2016). This has led authors to assume a unique orienting function for references placed within the Introduction. For example, Hargens (2000) restricted his citation context analysis to this section of the paper, to generate what he termed the "orienting reference list". However, location alone is not sufficient to provide an indication of how references differ across sections of the paper. Our results contribute to this discussion, by examining high frequency verbs by section. The strong presence of <Show> and <Suggest> in the Introduction and Discussion demonstrate an important rhetorical function of citations in these sections: providing demonstrative evidence upon which the current work builds. A shift to active stances is found in the Methods: instead of showing or suggesting, citations in the Methods take action: performing, calculating, and using. High frequency verbs, therefore, provide an indication of the functions of various references, based upon their location in the text.

Sentiment analysis of agreement and negation provided another window into the normative stance of citations. Our work demonstrated that negation was infrequent—reinforcing the citation studies demonstrating the low rate of negative citations (Catalini, Lacertera, & Oettl, 2015). Furthermore, agreement was most prevalent in the end of the results and beginning of the discussion: implying that the agreement demonstrated is between the cited document and the results of the citing document. This provides empirical evidence of the "bricklaying" that Price envisioned. It also has application for the construction of similarity indices for information retrieval purposes.

From a computational perspective, this work also contributes to creating dictionaries of cue words, particularly in demonstrating the diversity of 3-grams that compose certain high frequency verbs in scientific texts. This is illustrative of the analytic nuance that is necessary to fully extract verbs from these texts and contextualize in-text citations. There are several purposes for which this could have use: for example, to create summaries of content or identify similar papers for information retrieval purposes.

This is, however, more than a retrieval question. It is also a fundamental question for the sociology of science in that it reveals the relationship among works of scholarship and provides insight into how knowledge accumulates (Hargens, 2000). It may also contribute to more theoretically informed indicators. As Moed (2005) commented: *"Quantitative analysts of science could develop more 'qualitative' citation based indicators, along a contextual or a cognitive-relational viewpoint, thus abandoning the principle underlying most citation analyses that 'all citations are equal'. […] Contextual indicators are derived from the*

*passages in the full text of scholarly documents in which a particular document or set of documents is cited" (p. 130).*

*Limitations*

Of course, citation context analysis is always a matter of perspective. As Sula and Miller (2014) note, "everything is in the eye of the citer"—which may not align with the perspective of the reader (Willett, 2013). It may also suggest that citation context analyses that work for one discipline may not be applicable to another (Hyland, 1999). For example, philosophy exhibits more negativity overall than other studies fields (Sula & Miller, 2014) and differences have been observed in how different fields cite a single work (Chang, 2013; Cozzens, 1985), particularly in the humanities, which vary in both form and function of citations (Frost, 1979; Sula & Miller, 2014). Although the PLOS corpus is fairly generalist, it still tends toward the biomedical sciences. Therefore, further research is necessary that more fully analyses across disciplines.

Furthermore, there are limitations to automatic classification given the nuances of language. For example, some of the obtained patterns that belong to the same class carry different meanings (e.g., in Table 7, the n-gram *"did not find"* expresses negation). Another limitation of this approach is that it does not distinguish between the nature of citations, namely citations which are perfunctory, and other types of citations. While the classes that we have examined do not correspond strictly to such categories, they can be used as a starting point for the categorization of citation contexts. By analyzing all categories corresponding to a class of 3-grams, this approach can be considered for the task of ontology population, in which categories have to be assigned to citation contexts.

The establishment of formal links between the categories of citations and the 3-gram patterns is beyond the scope of this article although this work is a first step in this direction. In general, each 3-gram may appear in sentences that correspond to different categories, and it is not possible to establish a one to one correspondence between the 3-gram patterns and the categories. However, if we consider the most frequent 3-grams of each class, we can observe that these 3-grams express very similar meanings. For example, for the class <Observe> (see Table 2) we have 3-grams such as: *has been observed, have been observed, also observed in*, etc. We can make the hypothesis that these expressions appear for the most part in sentences that belong the categories such as "Cites for information" and "Confirms" from the CiTO ontology. For example:

- *'A similar phenomenon has been observed in mammalian orthologous homeotic complex genes [X].'* (category "Confirms")
- *'It has been observed that insertion of a transgenic selectable marker to make a gene knockout can influence the expression of neighbouring genes [X].'* (category "Cites for information")

Among the less frequent 3-grams for the class <Observe> we find the expression *"did not observe"* (see Table 7). This expression appears in sentences that belong, among others, to the category "Disagrees with". For example: *'We did not observe open conformations similar to those reported in crystal structures of other CYPs [X] or in a recent molecular dynamics study of soluble CYP2C9 [X].'* These examples show that the 3-gram patterns can appear in one or more categories.

## Conclusion

The purpose of this study was to demonstrate the existence of frequent n-gram patterns in citation contexts and their relation with the rhetorical structure of scientific articles. Studying the distribution of n-gram classes containing verb forms, we show the existence of three different types of distributions according to the rhetorical structure. We have seen that the use of the most frequent patterns in citation contexts is governed by the sections of the rhetorical structure of scientific articles. Studying such structures will lead to a better understanding of the various functions of citations. While we do not carry out a full semantic annotation of the citations, we propose a quantitative methodology which does not rely on external resources such as ontologies or linguistic resources. The limitations of this approach are related to the n-gram classes based on verbs. If they offer a first relevant classification, such classes are not sufficient to describe the complexity of the phenomena related to in-text citations.

This study on the n-gram classes gives us two important results. First, it shows that the rhetorical structure plays an important role in the distributions of the n-gram classes in texts and by extension raises the question of the relation between citation acts and this structure. Second, this approach allows us to identify sentences that can be potentially annotated with citation acts. From our point of view, the problem of the automatic annotation of citation contexts is strongly related to identifying of significant surface patterns for the annotation process. This methodology offers new perspectives to locate these discursive forms according to the rhetorical structure of scientific articles. Our future work will consist of implementing the automatic annotation of citation contexts by more linguistically motivated approaches. This will be a starting premise for future research on defining the framework of the study of acts of citations.

## Acknowledgments

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

## References

Athar, A. & Teufel, S. (2012). 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*: 597-601.

Bertin, M., & Atanassova, I. (2014). 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)*. Amsterdam, The Netherlands.

Bertin, M., Atanassova, I., Larivière, V., & Gingras, Y. (2013). The Distribution of References in Scientific Papers: an Analysis the IMRaD Structure. In *Proceedings of the 14th International Conference of the International Society for Scientometrics and Informetrics*. Vienna, Austria. 591–603..

Bertin, M., Atanassova, I., Gingras, Y., & Larivière, V. (2016). The invariant distribution of references in scientific articles. Journal of the Association for Information Science and Technology, 67(1), 164-177.

Bloch, J. (2010). A concordance-based study of the use of reporting verbs as rhetorical devices in academic papers *Journal of Writing Research, Citeseer, 2*, 219–244.

Boyack, K. W., Small, H., & Klavans, R. (2013). Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology*, *64*(9), 1759–1767.

Boyack, K. W., Klavans, R., Small, H., & Ungar, L. (2012). Characterizing emergence using a detailed micro-model of science: Investigating two hot topics in nanotechnology. In Proceedings of PICMET'12: Technology Management for Emerging Technologies. 2605-2611.

Bradshaw, S. (2003). Reference directed indexing: Redeeming relevance for subject search in citation indexes. In International Conference on Theory and Practice of Digital Libraries. Springer Berlin Heidelberg, 499—510.

Callahan, A., Hockema, S., & Eysenbach, G. (2010). Contextual cocitation: Augmenting cocitation analysis and its applications. Journal of the American Society for Information Science and Technology, 61(6), 1130-1143.

Catalini, C., Lacetera, N., & Oettl, A. (2015). The incidence and role of negative citations in science. Proceedings of the National Academy of Sciences of the United States of America, 112(45), 13823–13826.

Cavnar, W. B., &, Trenkle, J. M., & others. (1994). N-gram-based text categorization. Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval *Ann Arbor MI*, *48113*(2), 161–175.

Chang, Y.W. (2013). A comparison of citation contexts between natural sciences and social sciences and humanities. Scientometrics, 96(2), 535-553.

Chubin, D.E., & Moitra, S.D. (1975). Content analysis of references: Adjunct or alternative to citation counting? Social Studies of Science, 5, 423-441.

Cozzens, S.E. (1985). Comparing the sciences: Citation context analysis of papers from Neuropharmacology and the Sociology of Science. Social Studies of Science, 15, 127-153.

Cronin, B. (1981). The need for a theory of citing. *Journal of Documentation*, *37*(1), 16–24.

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

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

Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., & Radev, D. (2008). Blind men and elephants: What do citation summaries tell us about a research article?*Journal of the American Society for Information Science and Technology, 59*(1), 51–62.

Frost, C. (1979). The use of citations in literary research: Preliminary classification of citation functions. Library Quarterly, 49(4), 399-414.

Fujiwara, T., & Yamamoto, Y. (2015). Colil: a database and search service for citation contexts in the life sciences domain. Journal of Biomedical Semantics, 6, UNSP 38, 1—11.

Garfield, E. (1964). Can citation indexing be automated? In Statistical association methods for mechanized documentation, symposium proceedings. Washington, DC : National Bureau of Standards, Miscellaneous Publication 269. 189-192.

Garfield, E., & others. (1972). Citation analysis as a tool in journal evaluation. American Association for the Advancement of Science. Retrieved from http://www.elshami.com/Terms/I/impact_factor-Garfield.pdf.

Gilbert, G.N. (1977). Referencing as persuasion. Social Studies of Science, 7(1), 113-122.

Giles, C.L., Bollacker, K.D., & Lawrence, S. (1998). CiteSeer: An automatic citation indexing system. In E. Witten, R. Akseyn, & F.M. Shipman III (Ed.). Digital Libraries 98: The Third ACM Conference on Digital Libraries (pp. 89-98). New York: ACM Press.

Gipp, B., & Beel, J. (2009). *Identifying related documents for research paper recommender by CPA and COA*. Paper presented at the Proceedings of International Conference on Education and Information Technology, Berkeley.

Halevi, G., & Moed, H.F. (2013). The thematic and conceptual flow of disciplinary research: A citation context analysis of the Journal of Informetrics, 2007. Journal of the American Society for Information Science and Technology, 64(9), 1903-1913.

Hargens, L. L. (2000). Using the literature: reference networks, reference contexts, and the social structure of scholarship. *American Sociological Review*, 65(6):846-865.
Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. Applied linguistics, 20(3), 341-367.
Hopper, P. (1987). Emergent grammar. In Annual Meeting of the Berkeley Linguistics Society. Vol. 13, 139-157.
Jörg, B. (2008). Towards the nature of citations. In Poster Proceedings of the 5th International Conference on Formal Ontology in Information Systems (FOIS 2008)
Kaplan, N. (1965). The norms of citation behavior: Prolegomena to the footnote. American Documentation, 16(3), 179-184.
Leydesdorff, L. (1998). Theories of citation? *Scientometrics*, *43*(1), 5–25. doi:10.1007/BF02458391
Lipetz, B.A. (1965). Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. American Documentation, 16(2), 81-90.
Liu, M. (1993). Progress in Documentation the Complexities of Citation Practice: a Review of Citation Studies. Journal of Documentation, 49(4), 370–408.
Liu, S., & Chen, C. (2012). The proximity of co-citation.*Scientometrics, 91*(2), 495–511.
Liu, S., Chen, C., Ding, K., Wang, B., Xu, K., & Lin, Y. (2014). Literature retrieval based on citation context. Scientometrics, 101(2), 1293-1307.
Marici, S., Spaventi, J., Pavicic, L., & Pifat-Mrzljak, G. (1998). Citation context versus the frequency counts of citation histories. Journal of the American Society for Information Science, 49(6), 530-540.
Mei, Q., & Zhai, C. (2008). *Generating impact-based summaries for scientific literature.* Paper presented at the Proceedings of ACL '08, Columbus.
Merton, R. K. (1961). Singletons and multiples in scientific discovery: A chapter in the sociology of science. Proceedings of the American Philosophical Society, 105(5), 470-486.
Merton, R. K. (1957). Priorities in scientific discovery: a chapter in the sociology of science. American sociological review, 22(6), 635-659.
Moed, H. F. (2005). *Citation Analysis in Research Evaluation*, Information Science and Knowledge Management. Springer.
Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., Radev, D., & Zajic, D. (2009). *Using citations to generate surveys of scientific paradigms.* Paper presented at the Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Boulder.
Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, *5*(1), 86–92.
Nakov, P. I., Schwartz, A.S., & Hearst, M.A. (2004). *Citances: Citation sentences for semantic analysis of bioscience text.* Paper presented at the SIGIR 2004 Workshop on Search and Discovery in Bioinformatics, Sheffield.
Nanba, H., & Okumura, M. (1999). *Towards multi-paper summarization using reference information.* Paper presented at the The 16th International Joint Conference on Artificial Intelligence, Stockholm.
Nenkova, A., & McKeown, K. (2011) Automatic Summarization. In *Foundations and Trends in Information Retrieval* 5(2-3): 103-233.
O'Connor, J. (1982), Citing statements: Computer recognition and use to improve retrieval. Information Processing and Management, 18: 125–131.
Peroni, S., & Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17, 33-43.
de Solla Price, D. J. (1963). Little Science, Big Science. Columbia University Press. New York.
Sakita, T. I. (2002). Reporting discourse, tense, and cognition. Elsevier Science.
Schneider, J.W. (2006). Concept symbols revisited: Naming clusters by parsing and filtering of noun phrases from citation contexts of concept symbols. Scientometrics, 68(3), 573-593.
Siddharthan, A. & Teufel, S. (2007). Whose idea was this, and why does it matter? attributing scientific work to citations. In Proceedings of HLT-NAACL. 316-323.
Sieweke, J. (2014). Peirre Bourdieu in management and organization studies—A citation context analysis and discussion of contributions. Scandinavian Journal of Management, 30(4), 532-543.

Small, H. G. (2011). Interpreting maps of science using citation context sentiments: a preliminary investgation.*Scientometrics, 87*(2), 373–388.
Small, H. G. (2004). On the shoulders of Robert Merton: towards a normative theory of citation. Scientometrics, 60, 71–79.
Small, H. G. (1982). Citation context analysis. *Progress in Communication Sciences*, *3*, 287–310.
Small, H.G. (1979). Co-citation context analysis: relationship between bibliometric structure and knowledge. Proceedings of the American Society for Information Science, 16, 270-275.
Small, H. G. (1978). Cited documents as concept symbols. Social studies of science, 8(3), 327-340.
Spiegel-Rösing, I. (1977). Science studies: Bibliometric and content analysis. *Social Studies of Science, 7*, 97–113.
Sugimoto, C.R. (Ed.). (2016). *Theories of Informetrics and Scholarly Communication*. Berlin: De Gruyter Mouton, pp. 426.
Sula, C.A. & Miller, M. (2014). Citations, contexts, and humanistic discourse: Toward automatic extraction and classification. Literary and Linguistic Computing, 29(3), 452-464.
Swales, J. (1986). Citation analysis and discourse analysis. Applied linguistics, 7(1), 39-56.
Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. In : Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Sydney, Australia. 103—110.
Teufel, S., Siddharthan, A., & Tidhar, D. (2009, July). An annotation scheme for citation function. In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue. Association for Computational Linguistics, 80-87.
Voos, H. and Dagaev, K. (1976) Are all citations equal? or did we op cit your idem? *Journal of Academic Librarianship*, 6(1), 19-21.
White, H.D. (2004). Citation analysis and discourse analysis revisited. Applied Linguistics, 25(1), 89-116.
Willett, P. (2013). Readers' perceptions of authors' citation behavior. Journal of Documentation, 69(1), 145-156.
Zhao, D. & Strotmann, A. (2014). In-text author citation analysis: Feasibility, benefits, and limitations. Journal of the Association for Information Science and Technology, 65(11), 2348-2358.