STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Patterns in citation context: the case of the field of scientometrics

Wout S. Lamers[*], Nees Jan van Eck[*], Ludo Waltman[*] and Holger Hoos[**]

[*]*w.s.lamers@cwts.leidenuniv.nl; ecknjpvan@cwts.leidenuniv.nl; waltmanlr@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (the Netherlands)

[**]*h.h.hoos@liacs.leidenuniv.nl*
Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Leiden (the Netherlands)

Introduction

With full text of scientific publications increasingly becoming available in electronic formats, a wealth of information on the context of citation is now open to study. This full text context of citations opens up many new research opportunities in citation analysis and may help reveal fundamental properties of and patterns in the process of knowledge accumulation in the sciences. Of particular interest to us is to determine if systematic analysis of citation context can help shed light on what role previous literature plays when cited in new publications, how new authors use past literature to further their own arguments, and whether we can disentangle disciplinary modes of knowledge accumulation from more general archetypes of contributions made to the scientific landscape.

In this paper, we describe the results of our exploration of citation context for a subset of papers in the field of scientometrics. While not within the scope of this paper, our ultimate aim is to characterize publications by how they are used in subsequent publications. Before this is possible, a better understanding of the features of citation context and the way previous work is presented, as well as how to extract these features from the citation context, is required. Throughout this paper we will explore several ways of characterizing publications by their citation context, including two features introduced in previous work by Hyland (1999): the use of integral or non-integral structures when incorporating citations into sentences, and the use of reporting verbs.

Integral citations are those that use the cited author's name explicitly as part of the citing sentence, whereas non-integral citations occur in parentheses or take some other form such as a numbered reference or footnote (Swales, 1990). Choosing the latter style indicates an author's decision to emphasize the content of the cited work, whereas the former style emphasizes the writer of the cited work (Hyland, 1999). We suggest that integral citations may be used to preserve some distance between the citing author and the cited work, allowing a citing author to present another's work without implicitly endorsing it, while non-integral citation may favour generalizations and presenting another's work as accepted fact.

Verbs, meanwhile, contain information about the relationship between the citing article and the cited work (Bertin, Atanassova, Sugimoto, & Lariviere, 2016). We hope that verbs in citing sentences will prove specific enough to characterize the type of contribution made by the cited article, yet general enough to rise above discipline-specific use of language. Reporting verbs in particular seem useful candidate features to distinguish types of

contributions. These verbs have been well-studied for they are routinely used by authors not only to report on the actions performed in a cited work but also to convey the authors' attitude on those actions or the cited work. We follow Hyland (1999) in making a distinction between research verbs (procedural or physical acts associated with performing the research, e.g. *find*, *show*, *analyse*), cognitive verbs (mental processes, e.g. *consider*, *focus*, *understand*) and discourse verbs (requiring verbal expression, e.g. *discuss*, *state*, *suggest*) as subcategories of reporting verbs. Previous research into reporting verbs in academic publications has focused on describing their types and function (e.g. Thomas & Hawes, 1994; Thompson & Yiyun, 1991) or disciplinary differences (Hyland, 1999) and their change over time (Hyland & Jiang, 2017). Our goal is to leverage verbs to describe types of contributions to knowledge accumulation in research fields. In this paper, we will investigate the occurrence of reporting verbs in integral and non-integral citations and as function of time between the publication of the citing and the cited article.

Method

To build our corpus of citing sentences in the field of scientometrics, we first selected a set of scientometrics publications from Web of Science using the CWTS publication classification system (see Waltman & van Eck, 2012 for further details). This classification system clusters publications based on direct citation relations. We selected 13280 publications in the time period 2000-2016 that occurred in the cluster containing the publication introducing the *h-index* (Hirsch, 2005). These publications were subsequently matched against a database of Elsevier full text publication records (see Boyack, van Eck, Colavizza, & Waltman, 2018 for a description of the database). Using this database all Elsevier publications citing our set of scientometrics papers were retrieved at the citing sentence level. Of our scientometrics papers, 5688 were cited by publications in our full text database, producing a corpus of 39522 citing sentences. Duplicate sentences occur, as sentences may cite more than one of our scientometrics publications. These duplicates are retained since we are operating at the level of individual instances of citation to a particular paper within a sentence.

For each citation, we determined whether it is integral or non-integral by processing their accompanying sentence. First, the sentence was stripped of all citation labels. Subsequently, all characters between brackets were removed from the remaining character strings, leaving only the primary text of the citing sentence. We then searched this text for the last name of the first author of the cited paper. Citations were labelled integral if this author name was found, and non-integral if it was not.

The extraction of the verbs from each citing sentence required a different processing routine, implemented using the Natural Language Toolkit (NLTK) for Python (Bird, Klein, & Loper, 2009). This process included replacing all citation labels with neutral, single-word strings designed to be processed as proper nouns, followed by word tokenization and part-of-speech (POS) tagging. Using the obtained POS tags, all verbs were extracted from the sentences, excluding gerunds, present participles and past participles. This process of POS tagging in NLTK has a stated accuracy of 95%. After lemmatization of the remaining verbs, a document-term matrix was compiled using cited publications for documents and verbs as terms. Verbs with a total occurrence below 5 were discarded, and the 400 most frequent verbs were manually checked. Subsequently, a list of terms to exclude was compiled (consisting of *have* and *be*, which were altogether too frequent to be informative, and several terms mistakenly tagged as verbs) and an additional rule excluding the verb *cite* when followed by a noun, as these proved to be predominantly wrongly tagged instances of 'highly cited publications' and similar phrases. The process was repeated with these new rules in place to

arrive at a final verb matrix. Finally, lists of reporting verbs were compiled by taking them from the literature and by manually inspecting the most frequent verbs in our dataset.

Results

The corpus of citing sentences

As shown in Figure 1, the number of citing sentences mostly inceases over time. This skewness towards more recent years is to be expected since these are the publication years of citing sentences, which may cite any previously published scientometrics paper in our set. The initial years contain few citations, but Figure 2 shows that from 2005 onwards the ratio of integral versus non-integral citations rests fairly steadily between 20% and 30% of the citing sentences. In 2008 we find an outlier, during which 45% of citations are integral. This coincides with the establishment of the Journal of Informetrics (JoI) as the main source of citing sentences, supplying 44% of citing sentences in 2008, 63% of which use the integral structure. In subsequent years JoI continues to be the largest source of citing sentences, with on average 36% of its citations past 2008 using the integral structure.

Figure 1: absolute counts of integral and non-integral citations per year

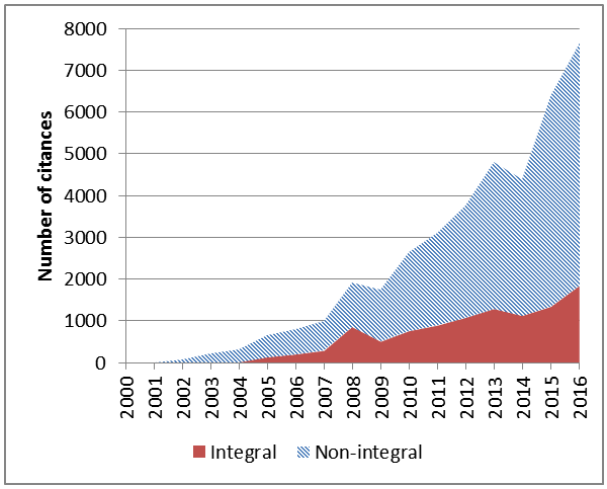
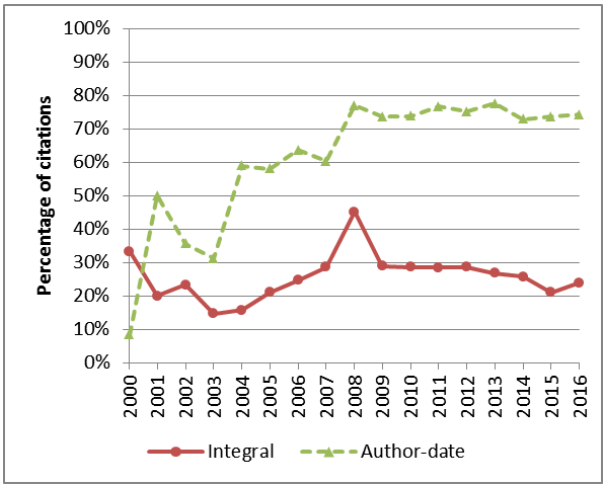


Figure 2: percentage of citing sentences adhering to listed citation styles



In earlier years, our corpus contains a roughly even mix of author-date style citations and other citation styles. From 2008 onwards, this balance moves in favour of author-date style citations, with about 75% of citing sentences containing years in their citation labels.

Location of citations in citing articles

For each citation it is possible to determine the location in the overall text progression of the citing article at the character level. This allows us to plot a curve of the location of citations along the text progression of the citing articles. Figure 3 displays these curves for all citations in the dataset, as well as for only integral and non-integral citations. Integral and non-integral citations occur in comparable positions in citing articles, with two notable exceptions: the very beginning, and the very end of articles. If one assumes these regions of papers are typically where one would make more general statements, this appears to support the idea that non-integral citations lend themselves to generalization, while integral citations may be used to present newer or more contentious work that the citing authors are not prepared to fully endorse.

Figure 4 shows the proportion of integral and non-integral citations in each twentieth of text progression. The dominance of non-integral citations at the beginning of papers is clear, while the disparity at the end of papers is less pronounced. Beyond these two extremes, integral citations occur at a relatively steady rate of 30% of citations.

Figure 3: Location of citations within citing articles

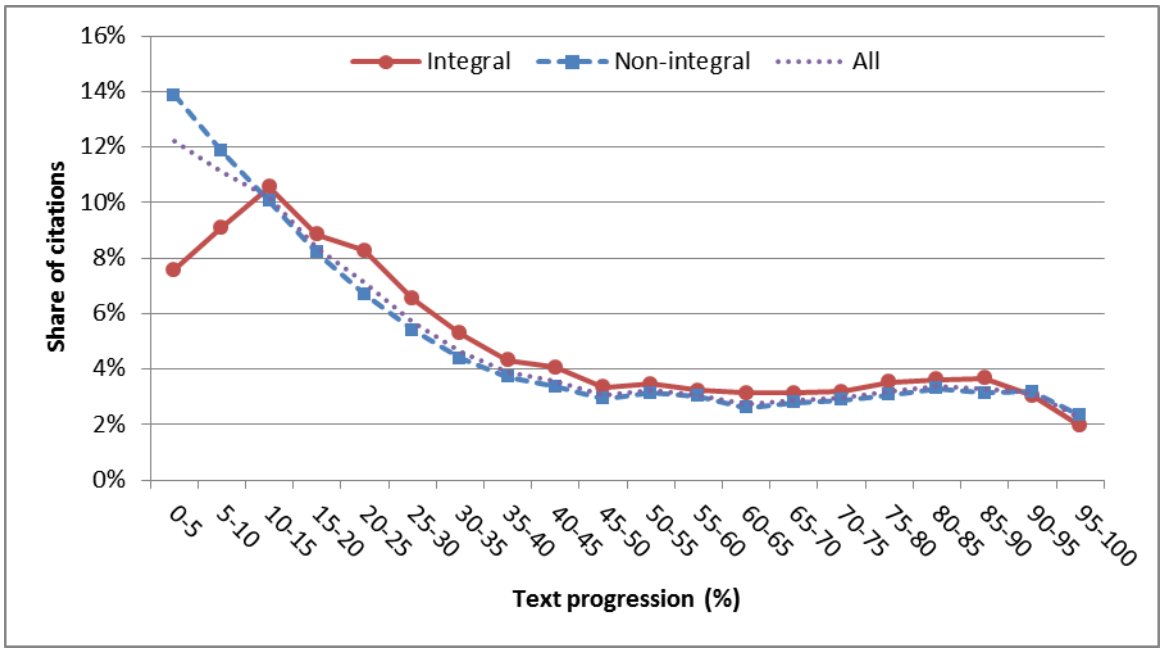
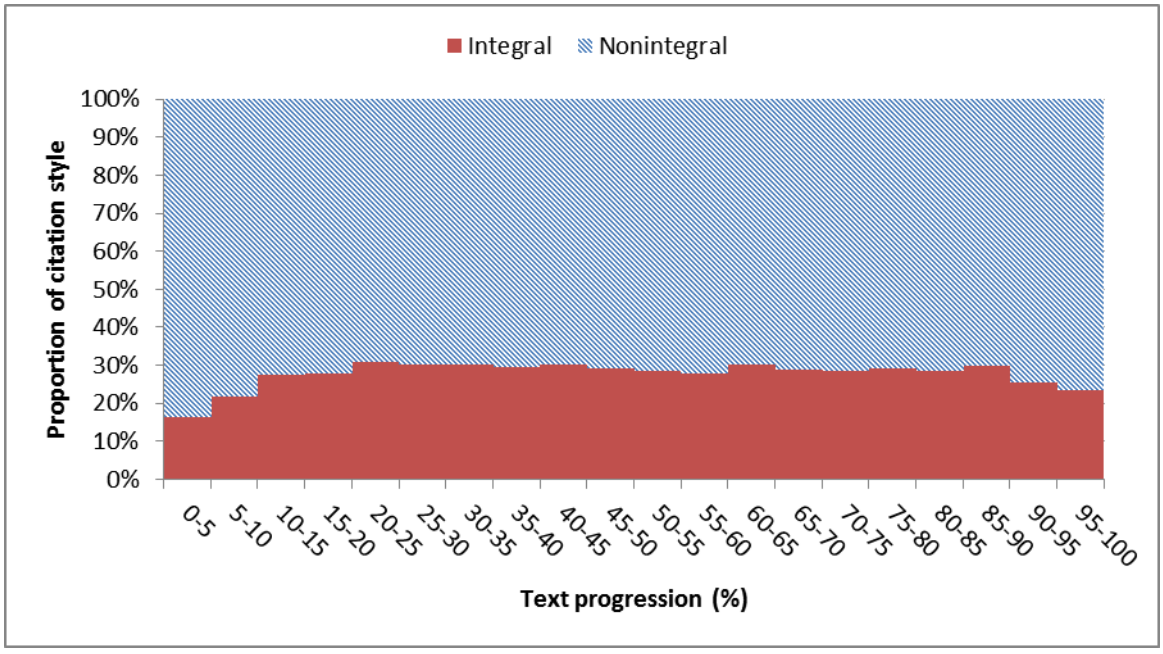


Figure 4: Proportion of integral and non-integral citations within citing articles



Reporting verbs

An overview of the most frequent reporting verbs by type, and their share occurring in integral sentences, can be found in Table 1. 38% of all verbs extracted from citing sentences were classified as some type of reporting verb. As Figure 5 shows, this percentage rises to 50% when only integral citations are considered, while non-integral citations contain

markedly fewer reporting verbs. Figure 6 makes clear that research verbs constitute just under 60% of reporting verbs, regardless of citation type. Integral citation sentences appear to contain more discourse verbs, compared to non-integral citations that contain comparatively more cognition verbs.

Table 1: Most frequent (reporting) verbs, their occurrence and its share in integral citations

Research verbs			**Cognition verbs**			**Discourse verbs**			**Non-reporting**		
Verb	*Count*	*% int*	*Verb*	*Count*	*% int*	*Verb*	*Count*	*% int*	*Verb*	*Count*	*% int*
find	1322	59%	consider	391	34%	suggest	650	41%	see	2000	82%
use	1311	43%	focus	372	25%	propose	485	72%	do	1345	73%
show	931	44%	reflect	324	15%	indicate	352	26%	include	1122	80%
identify	696	28%	explore	204	29%	report	330	38%	provide	1026	75%
analyze	532	52%	understand	157	26%	describe	269	40%	make	490	69%
study	478	57%	believe	77	22%	define	261	58%	increase	452	81%
measure	466	23%	think	75	19%	argue	238	56%	take	442	71%
evaluate	433	23%	recognize	74	26%	note	233	55%	follow	418	59%

Figure 5: Share of verb types

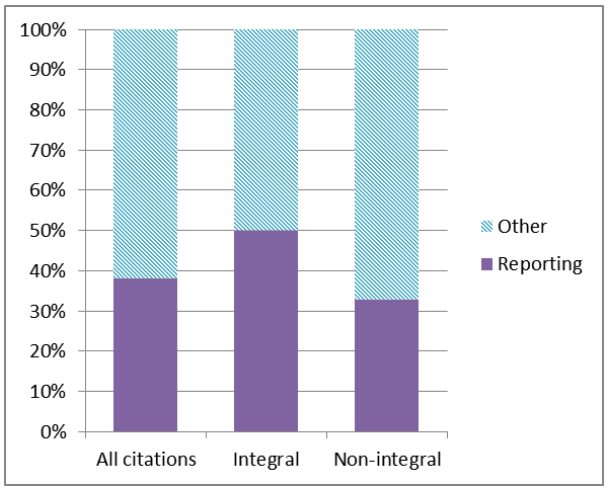
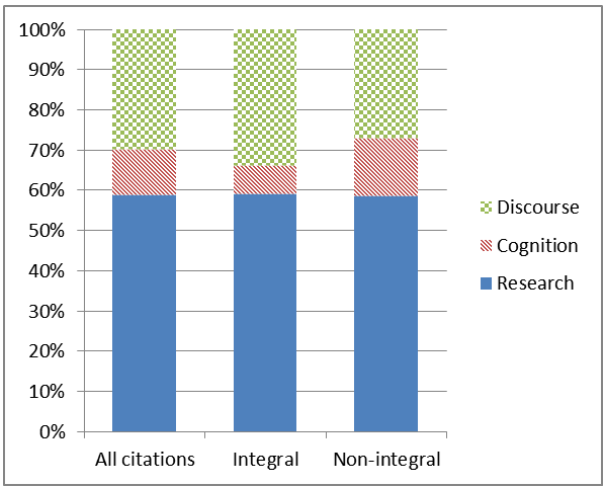


Figure 6: Types of reporting verbs



Time since publication of cited article

To investigate whether the usage of integral and non-integral forms, or types of reporting verbs, is influenced by the age of the cited publication, we plot these features against the number of years between citing and cited publications. As evident in Figure 7, the use of integral form gradually falls as the time between citing and cited publication increases. A possible explanation would be that as time goes on, publications are generally accepted into the canon of a research community, making authors more inclined to generalize their findings and more willing to accept the message presented in the work without needing to explicitly attribute it to another author within their narrative. On the other hand, Figure 8 shows little variation in the occurrence of the different categories of reporting verbs as time between citing and cited publication increases. The slightly different patterns at the end of the time scale can be attributed to the rapidly decreasing number of citations with a gap of 15 or more years between citing and cited publications.

Figure 7: Integral and non-integral citations and time since publication

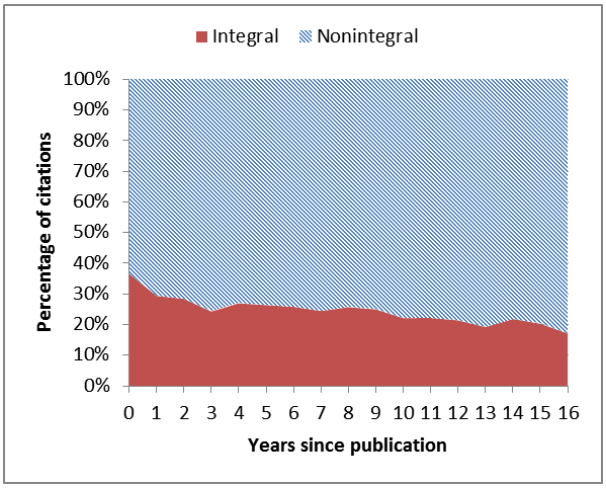
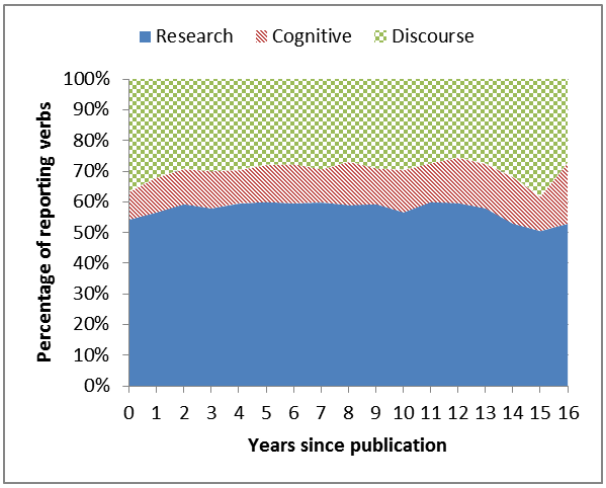


Figure 8: Reporting verb subclasses and time since publication



Association of verbs to integral and non-integral structures

Figure 5 has shown that reporting verbs occur relatively more in sentences with integral citations than in sentences with non-integral citations. Table 2 lists the 20 verbs for both integral and non-integral structures whose occurrences are most skewed towards one or the other, considering only verbs with an absolute occurrence of 50 or more. It is clear that this skewness extends to individual verbs, and that reporting verbs make up the majority of integral-specific verbs. *Encourage*, *know* and *reflect* are surprise entries in the list of verbs appearing most frequently alongside non-integral citations.

Table 2: Frequent verbs most strongly favouring sentences with (non-)integral citations

Verb	Count	% integral	Type
state	62	77%	Discourse
conclude	194	73%	Discourse
propose	485	72%	Discourse
discuss	134	69%	Discourse
introduce	214	69%	Discourse
observe	149	63%	Research
characterize	104	63%	Other
claim	60	60%	Discourse
find	1322	59%	Research
construct	78	59%	Other
define	261	58%	Discourse
present	218	57%	Discourse
study	478	57%	Research
argue	238	56%	Discourse
look	75	56%	Other
carry	95	56%	Other
note	233	55%	Discourse
prove	64	55%	Research
conduct	136	54%	Research
classify	85	53%	Research

Verb	Count	% non-integral	Type
encourage	82	95%	Discourse
lack	61	93%	Other
emerge	68	93%	Other
continue	130	92%	Other
play	123	91%	Other
learn	54	91%	Other
enhance	84	90%	Other
help	261	90%	Other
facilitate	92	89%	Other
ensure	73	89%	Other
know	70	89%	Cognition
bring	66	88%	Other
attract	90	88%	Other
divide	55	87%	Other
become	226	87%	Other
contribute	202	87%	Other
limit	59	86%	Other
consist	64	86%	Other
gain	88	85%	Other
reflect	324	85%	Cognition

Discussion

In this paper we set out to explore patterns in features of citing sentences, in particular their reliance on integral or non-integral structures and their use of reporting verbs. It has become clear that the distinction between integral and non-integral citations is a relevant one, with integral citations occurring markedly less frequently in the earliest parts of citing articles. Furthermore, we clearly find more reporting verbs in sentences with integral citations than with non-integral citations, and the use of integral structures steadily drops as the age of cited material increases. This appears to support our suggestion that integral structures might be used to present newer material that has not yet been firmly enshrined in the canon literature of a research field. Older publications, receiving mostly non-integral citations, might be more readily generalized by citing authors, as they have likely withstood the research community's scrutiny.

While reporting verbs appear to go hand-in-hand with integral citations, it is less straightforward to find patterns in their occurrence, in part due to their variety. Determining whether a verb qualifies as a reporting verb, let alone a specific subtype thereof, is debatable, and the exact meaning of a verb tends to be influenced heavily by its context. Nevertheless, it is clear that the occurrence of certain verbs favours certain types of citations and they remain a potent carrier of information regarding the citing authors' use and interpretation of the cited work.

Limitations of this study need to be noted. While our data set covered close to 40000 citing sentences, it was limited to a single field of research, which means that findings here are difficult to generalize. It has also proven challenging to perform analysis at the cited paper level because individual cited papers often only have a limited number of associated citing sentences. Of the 5688 unique cited scientometrics papers, only 264 had 25 or more associated citing sentences.

We intend to expand our analysis to other scientific fields in the future. Further research might also seek to leverage more sophisticated methods for the analysis of verbs in citing sentences, including using dependency parsing of citing sentences to find their core components or to find the verbs most closely associated with the cited literature in integral sentences. In addition to this, dimensionality reduction techniques such as topic modeling or the use of word embeddings might allow us to move beyond pre-established lists of reporting verbs and instead derive patterns in verb usage directly from the data. Perhaps such approaches may also enable us to disentangle discipline specific dimensions from more generic types of knowledge accumulation patterns.

References

Bertin, M., Atanassova, I., Sugimoto, C. R., & Lariviere, V. (2016). The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. *Scientometrics*, *109*(3), 1417–1434.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Boyack, K. W., van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics*, *12*(1), 59–73.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–72.

Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics*, *20*(3), 341–367.

Hyland, K., & Jiang, F. (Kevin). (2017). Points of Reference: Changing Patterns of Academic Citation. *Applied Linguistics*, (April), 1–23.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Thomas, S., & Hawes, T. P. (1994). Reporting verbs in medical journal articles. *English for Specific Purposes*, *13*(2), 129–148.

Thompson, G., & Yiyun, Y. (1991). Evaluation in the reporting verbs used in academic papers. *Applied Linguistics*, *12*(4), 365–382.

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392.