

# Towards the Nature of Citations

Brigitte Jörg

*German Research Center for Artificial Intelligence,  
Language Technology Lab, Saarbrücken, Germany  
brigitte.joerg@dfki.de*

**Abstract.** Citation Analysis has a long tradition in information science where it is a subfield of bibliometrics, now often called scientometrics or informetrics. Since its inception, when Eugene Garfield set up the Science Citation Index at the Institute for Scientific Information (ISI) in 1960, citation counts have matured towards a serious means of assessment for the impact of scholarly work. Along with their application grew the criticism about the quality of simple counts and various attempts were made to examine the citation content or context for getting deeper insight into the development and transfer of scientific knowledge. Several methods have since been employed for examination and classification of citations with respect to their function. With this paper we propose *cue verbs* (verbs within the citation function) to be highly supportive in understanding the citation relation and with it the knowledge acquisition process. Cue words have been categorized into classes, but they could be organized ontologically in multiple dimensions. A better understanding of the citation function will thus be the seed for a citation ontology, which we consider extremely useful for machine learning tasks, and in areas such as knowledge engineering and information system design.

**Keywords.** citation function, citation relation, cue verbs, classification, knowledge acquisition, knowledge engineering, conceptual modelling, formal ontology

## Introduction

Citations have not only been studied by information scientists. Sociologists of science have hoped to gain insight into the social construction of knowledge, and applied linguists have studied citers' behaviour to identify differences in disciplinary discourse and to teach citation in advanced writing courses. For quite some time, the three fields have been largely ignorant of each others work [14], and cross-fertilization has only started recently [18]. Serious attempts to resolve the problems with simple citation counts were first undertaken by Moravcsik and Murugesan [9] with an in-depth study of the organic nature of citations, by developing a typology to estimate their quality and context. In that same year, Chubin and Moitra [3] slightly modified this typology and suggested their categories to be mutually exclusive. Researchers from heterogeneous fields have since continued to invest into the study of references. Small [11] found a high degree of uniformity between concept symbols given in citations and the content of the cited documents in Chemistry. In more recent work, Teufel [15] has developed universal categories (argumentative zones: background, other, own, aim, textual, contrast, basis) to classify sentences in journal articles, showing that discourse studies are helpful for user- and task-tailored summaries and for improved citation analyses. Elkiss [4] found citing sentences to Biomedicine articles more focused than the cited document's abstract by comparing their inherent cohesion. Teufel [16] introduced a

citation annotation scheme, adapted from Spiegel-Rüsing [13], and inspired by the findings from Swales [14] that scientific argument follows a general rhetorical structure, to study the interplay of discourse structures of scientific arguments with formal citations.

The mentioned schemes require close reading, domain knowledge, and expert judgement to recover the implicit meaning of citations from text and therefore cannot easily be delegated to machines [18]. Different approaches were taken with analysing the syntax of sentences. For identification of relationship types, Pham and Hoffmann [8] developed rules with their KAFTAN tool from semi-supervised classification of sentences (basis, support, limitation, comparison) extracted from text by regular expressions, and supported by a list of synonyms taken from WordNet [5]. Nanba and Okumura [10] used cue words in reference information to understand relationships between papers for automated survey (a summary of multiple papers) generation.

We propose a different approach to get insight into the citation function and structure of scientific knowledge. First, we manually extract a huge number of *cue verbs* from text within the citation function; second, we align those extracted *cue verbs* according to their semantic similarity; before, third, we classify these aligned instances manually into a citation scheme comparable to [9] or [16]. These steps will direct us towards a citation ontology, and, in addition, will set the stage for supervised machine learning.

## 1. Citation Classification

The first approach towards a classification of citations proposed by Muravcsik and Murugesan [9] allows for multiple attributes. That means, for a single complex citation function, four values could be in use:

- conceptual / operational
- organic / perfunctory
- evolutionary / juxtapositional
- confirmative / negational

Teufel [16] defined her citation categories to be mutually exclusive:

- weakness of cited approach
- contrast/comparison in goals or methods (neutral)
- unfavourable contrast/comparison (current work is better than cited work)
- contrast between 2 cited methods
- author uses cited work as starting point
- author uses tools/algorithms/data
- author adapts or modifies tools/algorithms/data
- the citation is positive about approach or problem addressed (used to motivate work in current paper)
- author's work and cited work are similar
- author's work and cited work are compatible/provide support for each other
- neutral description of cited work, or not enough textual evidence for above categories or unlisted citation function

## 2. Citation Function and Cue Verbs

We consider as the citation content or context the text immediately around a reference. As a citation function we define the text area, embedding reference(s) to related work. Citation functions contain cue verbs such as the following:

- we **compare** ... **with** ...
- ... **lack**
- **same as** ..., we **adopt** ...
- ... **further explored**
- **for comparison** we **adopt** ...
- ... **adopted** ...
- ... **explore** ... in ...
- it is **reported** ... **that**
- **typical works include** ...

We assume the semantics of cue verbs to be constrained by the fact of their usage in citation functions, from where it is clear that a reference is made. We implicitly know that an author or a group refers either to a theory, another approach, a new method or technique. What we do not know is the nature of the relation itself: is it an agreement, is it a comparison, is it foundational, or negational. We consider the imposed verbal construction of citation functions crucial to understanding the nature of references, and we even assume the inherent semantics of cue verbs sufficient for an understanding of the citation function, its alignment and classification. Furthermore, we expect the number of referring cue verbs within citation functions to be finite, though there are many linguistic variations (i.e.: active, passive, tense, adverbials).

Our hypotheses are based on first thoughts and deduced from an examination of extracted cue verb examples, which suggest a classification scheme that allows for more than one value assigned to citation functions. Verification thereof will have to be achieved from multiple evaluators by tests of inter-annotator agreement. Our analysis process is composed of the following three steps:

- (1) **Extraction**      (2) **Alignment**      (3) **Classification**

## 3. Cue Verb Analysis within Citation Relations

We identified cue verbs within citation relations as important carriers of information. By investigation of 150 cue verbs in the ACL Anthology Network<sup>1</sup>, as referred to in [1], we detected similarity in their verbal construction as well as in their semantics within citation functions. Table 1 gives some insight into the nature of citations by classifying them (bold headings) according to the scheme in [9], which was built from an analysis of citation contexts in journal articles dealing with Theoretical Physics.

For the development of our citation ontology, we will examine cue verbs from articles in the highly dynamic and complex field of Language Technology<sup>2</sup>. We will

---

<sup>1</sup> ACL Anthology Network: <http://belobog.si.umich.edu/clair/anthology/index.cgi> enables online analysis over 14.000 papers in Computational Linguistics, collected from the ACL Anthology <http://www.aclweb.org/anthology-new/>.

<sup>2</sup> Language Technology World: <http://www.lt-world.org/>

use the ACL Anthology Network, which provides a public collection of noted articles in the range of Computational Linguistics and Language Technology, interlinked by citations, dating back to as early as 1965. For each article the incoming as well as the outgoing citations (within the network) are listed, and the citation summaries from incoming citations are available. These citation summaries contain the relevant text areas that include the cue verbs for our analysis, alignment and ontology development.

**Table 1.** Classification of aligned Cue Verbs in Citation Relations according to [9].

<b>organic/evolutionary/confirmative</b> based on ... in the sense of ...	<b>perfunctory/evolutionary/confirmative</b> introduced by ... proposed by ... described in ... described by ... in case of ... focussed around ... used in ...
<b>juxtapositional/negational</b> ... lack ... outperforms	<b>juxtapositional/confirmative</b> we compare with ... resembles ... are similar to ... same as ...
<b>perfunctory/evolutionary/confirmative</b> ... described ... proposed ... use ... employed ... describe ... employ a ... present ... focus on ... consider ... generalize ... explore	<b>operational/evolutionary/confirmative</b> were adopted from ... we build on this work ...
	<b>juxtapositional</b> differs from ...
<b>organic/evolutionary/confirmative</b> ... further explored ... have extended ... extend this work ... defines	<b>conceptual/evolutionary/confirmative</b> inspired by ... typical works include ... currently most works focused on ...

#### 4. Citation Ontology

We aim at getting more insight into the construction of knowledge as it is practiced in written scientific articles by citation relations. Articles contain units of knowledge that represent particular topics and belong to specific concepts. Relationships between the articles (their knowledge units or concepts) are explicitly composed as citation relations by bibliographic pointers and in natural language. This common citation system does not comply with any formal means or semantics that would allow for a large-scale analysis of the relationships between the articles, and consequently, the construction of knowledge cannot easily be followed. We propose cue verbs within citation relations to guide our steps towards a citation ontology that will allow us to access the structure in scientific knowledge.

Table 1 shows important functions of citation relations inherent in cue verbs, deduced from articles in theoretical physics [9]: *operational*, *conceptual*, *organic*, *perfunctory*, *confirmative*, *negational*, *evolutionary*, *juxtapositional*. To understand the development of knowledge, our citation ontology may need further views: *foundational* (inspired by), *state-of-the-art* (currently most works focused on, typical works include), *experimental* (explore, have extended). Additionally, models of discourse relations

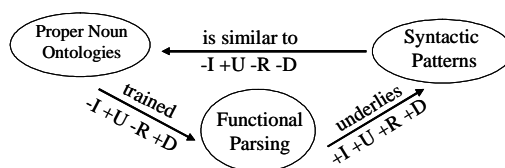
*supports, contradicts, suggests, suggests that not, proves, refutes, agrees, disagree* [17] or coherence relations [12] seem relevant for a citation ontology.

The systematic approach towards a citation ontology requires an ontological analysis of citation functions and their cue verbs. Guarino and Welty [6] (OntoClean)<sup>3</sup> investigated taxonomic relations by defining a set of meta-properties (rigidity, dependence, identity, unity) to clarify the arguments behind taxonomic relations (is-a, class inclusion, subsumption). Table 2 shows an example of their meta-properties [6] as imposed on citation functions and cue verbs. These meta-properties will not only be highly valuable for the study of arguments behind citation relations, but may also aid in delimiting conceptual or knowledge units.

**Table 2.** Meta-Properties [6] assigned to Citation Functions [9] and Cue Verbs.

Citation Functions / Cue Verbs	Meta-Properties
<b>juxtapositional/confirmative</b> we compare with ...	carries unity <b>+U</b> , does not carry identity <b>-I</b> , is not dependent <b>-D</b> , is not rigid <b>-R</b>
<b>conceptual/evolutionary/confirmative</b> inspired by ...	carries unity <b>+U</b> , supplies identity <b>+O</b> , is dependent <b>+D</b> , is rigid <b>+R</b>
<b>organic/evolutionary/confirmative</b> ... have extended based on ...	carries unity <b>+U</b> , carries identity <b>+I</b> , is dependent <b>+D</b> , is rigid <b>+R</b>
<b>perfunctory/evolutionary/confirmative</b> ... described ... proposed	carries unity <b>+U</b> , carries identity <b>+I</b> , is not dependent <b>-D</b> , is not rigid <b>-R</b>

We have concentrated on the reference semantics between citing and cited papers based on cue verbs. However, cue verbs do not provide domain information. For a contextual evaluation and validation, citation functions have to be re-constructed either manually, or automatically e.g. with tools for automated topic extraction [2]. Entire citation relations may be re-composed as triples  $\langle \text{Topic}, \text{Function}, \text{Topic} \rangle$  to build overarching knowledge maps or domain ontologies enhanced with meta-properties as in figure 1.



**Figure 1.** Example Knowledge Map enhanced with Meta-Properties

## 5. Results and Discussion

From the study of cue verbs and by inter-annotator agreement, a decision about the exclusiveness of the categories has to be taken: should categories be mutually exclusive or do multiple values support our understanding. Our preliminary study shows that some functional values (i.e. confirmative) depend on others (i.e. organic), what may be

<sup>3</sup> OntoClean: a methodology for ontology-driven conceptual analysis:  
<http://www.ontoclean.org/>

defined as a rule. For a complete picture or map, the papers not cited also need to be investigated. Finally, from our analysis and citation ontology we expect to get inspiration for engineering and ontology-driven system design that actively takes into account the citation function during the authoring process as requested by [17]. Litman [7] found cue phrases to be very effective for the improvement of machine learning algorithms in their classification tasks. The automated detection of relevant sentences and the disambiguation of relevant cue verbs within citation functions can be supported by dependency parsers such as MINIPAR<sup>4</sup>. A large set of cue verbs, annotated with properties from a citation ontology will be a valuable asset for ontology learning and topic extraction [2] and for an automated re-construction of knowledge across domains.

## Acknowledgement

I wish to thank Prof. Hans Uszkoreit for very valuable comments and discussions.

## References

- [1] Bird, S.; Dale, R.; Dorr, B. J.; Gibson, B.; Joseph, M. T.; Kan M-Y.; Lee, D.; Powley, B.; Radev, D.R. Tan, Y.F.: The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. InProceedings: *LREC 2008*, Marrakesh, Morocco, May 2008.
- [2] Buitelaar, P. and Eigner T.: Topic Extraction from Scientific Literature for Competency Management. InProceedings: *PICKME workshop on Personal Identification and Collaborations: Knowledge Mediation and Extraction*. ISWC 2008, Karlsruhe.
- [3] Chubin, D. E. and Moitra S. D.: Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science* 5(4): 423-441, 1975.
- [4] Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., Radev, D.: Blind men and elephants: What do citation summaries tell us about a research article? *JASIST Journal*, Vol. 59, No. 1, 2008, 51-62. John Wiley & Sons, Inc., New York, USA.
- [5] Fellbaum, C.: Wordnet – an electronic lexical database. *MIT Press, Cambridge, MA*, 1998. 764, 765.
- [6] Guarino N., Welty, C.: Ontological Analysis of Taxonomic Relationships. *LNC3 2000*. 1920, pp 210-224, 2000.
- [7] Litman, D. J.: Cue Phrase Classification Using Machine Learning. *Journal of Artificial Intelligence Research*. Vol. 5, 53–94, 1996.
- [8] Pham, S. B. and Hoffmann, A.: A new Approach for Scientific Citation Classification Using Cue Phrases. *Lecture Notes Artificial Intelligence 2003*, LNAI 2903, pp. 759-771; 2003.
- [9] Muravcsik M. J. and Murugesan, P.: Some results on the function and quality of citations. *Social Studies of Science*, 5:88-91, 1975.
- [10] Nanba, H., and Okumra, M.: Towards multiple-paper summarization using reference information. In Proceedings: *International Joint Conference on Artificial Intelligence*, Stockholm Sweden, 762–770, 1999.
- [11] Small, H. G.: Cited Documents as Concept Symbols. *Social Studies of Science*. 8:327-340, 1978.
- [12] Sanders, T.; Spooren W.; Noordman, L.: Towards a Taxonomy of Coherence Relations. *Discourse Processes*. 15:1-35 (1992).
- [13] Spiegel-Rüsing, I.: Bibliometrics and content analysis. *Applied Linguistics*. 7(1):39-56, 1977.
- [14] Swales, J.: Citation Analysis and Discourse Analysis. *Applied Linguistics*. 7(1):39-56, 1986.
- [15] Teufel, S.: Argumentative Zoning: Information Extraction from Scientific Text. *PhD Thesis*. School of Cognitive Science, University of Edinburgh, UK.
- [16] Teufel, S.; Siddharthan, A.; Tidhar, D.: An annotation scheme for citation function. In Proceedings: *7<sup>th</sup> SIGdial Workshop on Discourse and Dialogue*, pages 80-87, Sydney, July 2006, ACL.
- [17] Waard, A. de; Breure, L.; Kircz, J.G.; Oostendorp, H. van: Modeling Rhetoric in Scientific Publications. *Current Research in Information Sciences and Technologies*. pp. 352-356, 2006.
- [18] White, H. D.: Citation Analysis and Discourse Analysis Revisited. *Applied Linguistics*. 25/1:89-116, Oxford University Press, 2004.

---

<sup>4</sup> MINIPAR is a broad-coverage parser for the English language:  
<http://www.cs.ualberta.ca/~lindek/minipar.htm>