A descriptive analysis of cite units from the perspectives of content and linguistic expressions

Watanabe, Koichiro; Kageura, Kyo; Sekine, Satoshi

2021

Watanabe, K., Kageura, K. & Sekine, S. (2021). A descriptive analysis of cite units from the perspectives of content and linguistic expressions. Library and Information Science Research E-Journal, 31(2), 104-118. https://dx.doi.org/10.32655/LIBRES.2021.2.2

https://hdl.handle.net/10356/155404

https://doi.org/10.32655/LIBRES.2021.2.2

volume 31, issue 2, pages 104-118 (2021). https://doi.org/10.32655/LIBRES.2021.2.2

A Descriptive Analysis of Cite Units from the Perspectives of Content and Linguistic Expressions

Koichiro Watanabe
Graduate School of Education, The University of Tokyo, Japan
Center for Advanced Intelligence Project, RIKEN
kouichirou-watanabe495@g.ecc.u-tokyo.ac.jp

Kyo Kageura
Graduate School of Education,
The University of Tokyo, Japan
kyo@p.u-tokyo.ac.jp

Satoshi Sekine
Center for Advanced Intelligence Project, RIKEN,
Tokyo, Japan
satoshi.sekine@riken.jp

ABSTRACT

Background. While the use of citations for assessing research impact is well-studied, there is little work that investigates the content introduced into the citing documents through citations and the linguistic expressions used to represent the cited content

Objectives. This study analysed the types of content introduced into citing documents using the citations (cited content) and units of linguistic expressions used to represent the cited content.

Methods. We classified the expressions representing the cited content according to the unit of linguistic expressions (terms and clauses) and classified the cited content into conceptual categories. We adopted different frameworks for the classification of cited content represented by terms and clauses. The categories for terms were domain specific and the categories for clauses took into account subjectivity and generality. We also described the detailed categories of cited content with examples and provided seven types of cited content for clauses.

Results. We found that among the expressions representing cited content, terms constituted about 40% and clauses constituted about 60%. The majority of the cite terms were domain specific; and 35% of the cite terms referred to unique concepts. Of the cite clauses, 50% were objective and term-specific, 40% were objective and general, and 10% were subjective.

Contributions. This research provided a description of cite units and elaborated on the categories for cited content. The results showed basic types of cited content and clarified the distribution of cite units.

Published by WKW School of Communication & Information & NTU Libraries,
Nanyang Technological University

INTRODUCTION

This study investigated the content introduced into a citing document[1] by using citations. Reits (2004) defined a citation as: "A written reference to a specific work or portion of a work (book, article, dissertation, report, musical composition, etc.) by a particular author, editor, composer, etc. that clearly identifies the document in which the work is to be found" (Reits, 2004, p. 142). Citations in scientific papers are identified by citation anchors. An example is:

> [1] *Classical Arabic has some 9000 roots, of which 1200 roots are in MSA (Habib, 2008).*

In this example, the citation anchor is "(Habib, 2008)."

In this paper, we refer to what is introduced into citing documents by using citations as a cite unit. A cite unit has two aspects: the linguistic expressions (cite expressions) and content (cited content). Double quotation marks ("") will be used to indicate cite expressions of cite units and angle brackets (<>) for cited content. When we refer to cite units that cover these two aspects or without distinguishing them, double quotation marks will be used for convenience. The context will disambiguate the usage of double quotations.

Small (1978) pointed out that what a researcher actually cites is a concept or an idea. This concept or idea corresponds to cited content in our study. Below is an example from Small:

> [2] *Small's paper (1) presents a completely erroneous interpretation of citation practice.*

In this example, the cite unit (expression and content) is "a completely erroneous interpretation of citation practice." This cited content is the idea attributed to the cited document. In the example [1], the cite expression representing the cited content is a sentence. Cite expressions and cited content are explained in detail in the Collecting Cite Units and Data section.

Small collected cite units from scientific articles and analysed how the documents were cited. Following Small's proposal that the cited content is a concept or an idea, several researchers analysed document collections (Schneider, 2006) and measured the research impact of scientific articles based on the spread of a concept or an idea (Bornmann, Wray, & Haunschild., 2019). The method used in these studies is called citation concept analysis (Bornmann, Wray, & Haunschild, 2019). These studies focused on domain specific noun phrases, such as "hydrogen scattering factor" (cf. Small, 1978).

While fully appreciating the importance of citation concept analysis, we identify two areas that need further exploration. The first concerns the cite expressions. While existing research focused on terms, expressions of cite units can be clauses, as is shown in example [1]. The second concerns the types of cited content. Although research so far has focused on domain specific cited content, non-domain specific content can appear as cited content. We expect that descriptive explorations in these two directions will widen the scope of citation concept analysis and contribute to further clarifying the concept and act of citation.

To clarify the unit of cite units from the perspective of linguistic expressions and their distributions as well as the types of cited content, we collect cite units, classify cite units according to the types of cite units from the perspective of language expressions and the content. We also observe individual instances of cite units. In collecting cite units, terms and

[1] A document which cites other documents is called a citing document, and documents which are cited are called cited document(s).

clauses are identified as the unit of cite expressions. In the classification of cited content, we use categories that fit the attribute of scientific papers. While some studies created conceptual categories for the linguistic expressions, the conceptual categories established specifically for cited content do not exist, to the best of our knowledge.

This study has two major contributions: clarification of how cite units are distributed from the perspective of linguistic expressions and content, and creation of specific categories for cited content at the clause level. While many conceptual categories have been proposed, little work has been done to derive categories for content represented by clauses. In the detailed analysis about the types of the cite clauses, we conducted categorisation of instances and showed the more elaborated types of cited content with the observation on the examples of the cite clauses. These results descriptively clarify the types of knowledge conveyed from a scientific article to another scientific article because citations connect knowledge claim[2] to knowledge which has been previously accepted (Leydesdorff, 1987), and provide basic information for the act of citing.

This paper is organised as follows. In Related Work section, we review previous studies which categorised citation contexts. Citation contexts are the linguistic expressions around a cite unit (Small, 2011). In the Method and Data section, we collect and classify cite units extracted from the data using the framework. In the Result and Analysis section, we present the result of the analysis and classification of cite units. In the Conclusion section, we summarise the results of this research and propose future work.

LITERATURE REVIEW

Several studies in the field of natural language processing have classified citation contexts, focusing on the functions and the roles of citations (Cohan, Ammar, Zuylen, & Cady, 2019; Jurgens, Kumar, Hoover, McFarland, & Jurafsky., 2016; Li, He, Meyers, & Grishman, 2012; Lin, 2018; Moravcsik & Murugesan, 1975; Taşkin & Al, 2018; Teufel, Siddharthan, & Tidhar, 2006; Widyantoro, Khodra, Trilaksono, & Aziz, 2013). This classification is called citation function classification (Abdallah, Zhendong, Tarus, & Arshad, 2019; Hernádez-Alvarez & Gomez, 2016). Although these studies have classified the linguistic expressions related to citations, the problem statement is different from that of our study, in that these have categorised citations from the point of view of functions or roles within the citing documents. For instance, Cohan, Ammar, Zuylen and Cady (2019) derived the categories “Background”, “Method”, and “Result comparison” to characterise citation contexts. These categories are not intended to be applied to the cited content itself. Some studies took into account the sentiment of authors in a citing document towards the cited content (Li, He, Meyers, & Grishman 2012). The sentiment reflects judgements of the citing author, but is not a direct attribute of the cited content.

Valenzuela, Ha and Etzioni (2015) and Wang et al. (2020) focused on the importance of cited content within a citing document. The importance is also decided by the attitude of the citing authors as well as the function of the citation context.

Kang and Kim (2012) categorised citation contexts according to the syntactic type of the expression, such as “phrase,” “clause,” “sentence,” “multi sentence,” and “others,” and described their distribution. They found that 10.6% of citation contexts are phrases, 7.8% are clauses, 75.2% are sentences, 5.2% are multi-sentences, and 1.2% are others. Färber and

[2] Leydesdorff (1987) contrasted knowledge claims and accepted knowledge. Knowledge claims are claims proposed as new knowledge in scientific papers.

Sampath (2008) found five types of citation contexts from the perspective of grammatical functions: “in-text,” “author,” “concept,” “claim,” and “incomplete.” These studies focused on linguistic expressions of citation contexts but not the content of the citation contexts. As shown in this section, the categorisation of citation contexts focused on the attitude of the citing authors or the linguistic expression used to represent the citation context. Researchers have focused on the role the cited content plays in the citing document, but have not analysed the linguistic expression representing cite units and the types of cited content.

METHOD AND DATA

In this section, we explain the procedures of our analysis. Firstly, we describe how we collect cite units. We elaborate on the procedure with examples. Secondly, we introduce two frameworks for the classification of cited content. As we will see, we introduce different frameworks of the conceptual categories for terms and clauses because the expressions of cite units are linguistically categorised into terms and clauses and these bear the different types of content.

Collecting Cite Units and Data

We assume that a sentence containing a citation anchor also contains a cite unit. We identify terms[3] and clauses as cite expressions in our analysis. For simplicity, we shall refer to the expression of a cite unit at the term level as a cite term and to the expression of a cite unit at the clause level as a cite clause.

We collect sentences containing a citation anchor and identify the cite unit from the perspective of linguistic expression. When a citation is for a term, we extract the term and regard it as a cite term. Below is an example of a cite term:

> *Our method is based on the Stanford Core NLP tools (Stanford Core NLP Tools, 2016) and MetaMap tool (Aronson, 2006) in UMLS.*

In this example, “(Stanford Core NLP Tools, 2016)” and “(Aronson, 2006)” are the citation anchors, and the cited content is <the Stanford Core NLP tools> and <MetaMap tool>. This whole sentence describes the methodology in the citing document, and the cited content are shown as the tools adopted in the citing document. In this example, the expression of the cite unit is a term.

Although the whole sentence that carries a citation anchor is sometimes regarded as a citation sentence (Bonab, Zamani, Learned-Miller, & Allan., 2019), the actual cite unit may be only a part of the sentence. In Example [2] in the Introduction, Small did not regard “Small’s paper (1) presents” as the expression of a cite unit. As in this example, when the predicate of a clause indicates the act of finding, proposing or showing (e.g., “propose”, “show”, and “find”), only its object4 is regarded as the expression of a cite unit. An example is:

> *Al-Khatib (2016) proposes actions to protect inexperienced authors against predatory journals.*

[3] Terms represent what we may roughly call concepts and knowledge referents.

In this example, the part “actions to protect inexperienced authors against predatory journals” is regarded as the expression of the cite unit. This expression of the cite unit is also at the term level.

When the whole sentence with a citation anchor contains one clause and can be regarded as the expression of a cite unit, the sentence itself is identified as the expression of a cite unit.

> *Any ambiguity at the morphological level will have an impact on the syntactic level (Attia, 2008).*

In the above example, the whole sentence is the expression of the cite unit. In the citation in this sentence, we consider the expression of the cite unit to be at the clause level.

We also consider the expression of the cite unit below to be at the clause level:

> *Xu, Liu and Fang (2011) found that, in contrast to other broad domain categories, OA articles in the humanities are at a disadvantage.*

In this example, the cited content is the object of “found”, that is, <that, in contrast to other broad domain categories, OA articles in the humanities are at a disadvantage>.

When a sentence with citation anchors contains multiple clauses, we divide the sentence into clauses and regard the clauses that represent cited content as the expressions of cite units. An example is:

> *Though the field of automatic text summarization is over half a century old (Luhn, 1958), there are many problems awaiting effective solutions.*

In this example, the expression representing the cited content indicated with “(Luhn, 1958)” is “the field of automatic text summarization is over half a century old”. This expression is not taken verbatim from Luhn[4]. The content of < the field of automatic text summarization is over half a century old > is evidenced by the document indicated with “(Luhn, 1958)”.

The procedure to collect cite units is as follows:

1. Collect sentences containing a citation anchor
2. Identify expressions of cite units
3. Extract cite units
    - 3.1. If an expression of a cite unit is at the term level
        - 3.1.1. Extract the cite term
    - 3.2. If an expression of a cite unit is at the clause level
        - 3.2.1. Divide the sentence to clauses
        - 3.2.2. Extract the cite clause(s)

We investigated scientific papers published in 2018 in the field of library and information science, from the following journals: *Information Processing and Management, Journal of the Association for Information Science and Technology*, and *Scientometrics*. These journals were selected intentionally because we have background knowledge of the topics they cover, which we expect facilitate our analysis. We chose 30 papers randomly from a set of all research papers published in 2018. From body text (not including abstract section) of these papers, we extracted sentences with citation anchors5 and collected altogether 1,614 cite units.

[4] If the expression was written by Luhn (1958), this would be regarded as a plagiarism.

[5] We extract only sentences with citation anchors and do not consider the sentence following the sentences with citation anchors.

Conceptual Categories

As terms and clauses represent different types of content, we use different classification frameworks for cite terms and cite clauses.

Classification Framework for Cite Terms

Many conceptual categories have been proposed to characterise the concepts represented by terms. Some conceptual categories are domain independent (Douglas, 1998; Gangemi, Guarino, Masolo, Oltramari, & Schneider, 2002; Giancarlo & Gerd, 2010; Grenon, Grenon, Smith, & Goldberg, 2004; Herre et al., 2001; Kitamura, Sano, Namba, & Mizoguchi 2002; Niles & Pease, 2001; Terziev & Manov 2005), while others are domain specific (Alomari, 2016; Anderssone & Gronlund, 2009; Elizarova, Khaydarov, & Lipachev, 2017; Ihsan & Qadir, 2019; Kalemi & Martiri, 2011; Ko, Song, & Lee., 2016; Pertsas & Constantopoulos, 2016; Shum, Motta, & Domingue, 2000).

We adopted the conceptual categories proposed by Ko, Song and Lee, (2016), as they were established as an ontology of the terms in scientific papers.

As mentioned in the Introduction, previous work in citation concept analysis has focused on domain specific noun phrases, which we also follow in this study.

We modified the categories by Ko, Song and Lee. (2016) to take into account concepts specific to library and information science. We do not observe the cite terms classified into Y. Instance in the framework by Ko, Song and Lee. (2016), which is a category for names, because of the nature of this domain. The category of Y. Instance is deleted in our study. While we recognise that it is important to consider whether a cite term refers to a concrete individual or item, no cite term was found to mention a name. We classified the cite terms based on Ko, Song and Lee (2016), and whether a cite term is a proper noun or a common noun.

Classification Framework for Cite Clauses

Some frameworks of conceptual categories for clauses have been proposed. Halliday (1994) and Dik (1997) proposed the frameworks focusing on the attribute of predicates in a clause, such as tense within the theory of Functional Grammar. Smith (2003) proposed Situation Entity, which focuses on the subjectivity and generality of content represented by a clause. Asher, Benamara and Mathieu (2008) proposed the framework for the classification of opinions represented by clauses.

While all of the frameworks for the classification of clauses focus on the predicate of a clause, which includes its tense, perfective, and modality, only Situation Entity considers the generality of content represented by clauses. Holmes (1987) stated that scientific papers mainly reports the investigation and ideas of researchers.[6] The former describes past events and the latter presents as general descriptions. As Holmes (1987) stated, generality of content is important to scientific knowledge. The difference in generality can influence roles in reasoning. An individual experience cannot directly evidence an idea. Based on this understanding, distinguishing between past events and general descriptions is important in scientific papers. Following up this idea, Situation Entity is adopted as our framework for classification of cite clauses.

[6] Holmes (1987) compared the “generalized observation” and “particular experience” with citing Dear’s work. Hence events and general descriptions can be thought to be contrast.

Situation Entity is composed of three categories at the first level:

- *Eventuality*: The category of *Eventuality* is of a clause describing a phenomenon appearing during a particular period, which represents term-specific and objective content.
- *General stative*: The category of *General stative* is of a clause describing a general phenomenon regardless of time. This is the category for the clauses that represent general and objective content.
- *Abstract entity*: The category of *Abstract entity* is of a clause describing someone's thought which is not physical and objective. This is the category for clauses that represent subjective content.

It has been pointed out that it is difficult to detect content of clauses into *Abstract entity*, because the classification cannot be conducted linguistically (Friedrich & Palmer, 2014). In fact, we can see the diversity of the criteria and the method for the classification based on Situation Entity (Friedrich & Palmer, 2014; Palmer, Ponvert, Baldridge, & Smith, 2007). Although the classification of content into *Abstract entity* is difficult, this is essential since the attribute of the *Abstract entity* is different from the attribution of the other categories. To address this difficulty, we classified the clauses which represent the subjective content with modality or some adjectives such as "significant" into *Abstract entity*. This is based on the idea that *Abstract entity* is the category for clauses expressing what does not exist in the physical world.

Each category consists of two sub-categories, but we do not use the sub-categories because we focus on the subjectivity and the generality of clauses in this study, following Holmes (1987).

RESULT AND ANALYSIS

As stated in the Collecting Cite Units and Data section, we identified 1,614 cite units, and these can be divided into terms and clauses according to the linguistic expressions. The cite units comprise 638 terms (39.5%) and 976 clauses (60.4%). The number of cite units taking the clausal form is about 1.5 times more than those taking term form.

In this section, we show the distribution of the cited content through a quantitative analysis, and present examples in each category.

Analysis of Cite Terms

Overall Tendencies of the Cite Terms

Table 1 summarises the result of the classification of cite terms. Our classification framework for cite terms contain six major categories, each with sub-categories. Two points can be noted in Table 1. The first point is that the percentage of domain specific terms is relatively high; the second point is that the percentage of proper nouns is less than 40%.

Examining the percentages listed in column 1 of the table, the categories of *A. Object*, *D. Theory Method*, and *E. Format Framework* comprise about 80% at the first level of the classification. The percentage of *D. Theory Method* is more than 40%. The percentage for each of the categories *B. Action Function*, *C. Property*, and *X. General Common* does not reach 10%. We find that the cite terms classified as *D. Theory Method* and *E. Format Framework* are domain specific, such as <the Gate NLP framework> and <a method based on the online ontologies of entities and aspects to summarize opinions>. This indicates that many citations at the term level are for introducing domain specific (i.e., technical) terms.

Table 1. The result of the classification of cite terms

Category	%	n	% (PN)	n (PN)
A. Object				
A01. Human	0.0	0	0.0	0
A02. Institution Organization	0.3	2	100.0	2
A03. Natural Object	0.0	0	0.0	0
A04. Artifacts	22.7	140	49.2	69
B. Action Function				
B01. Action Activity Role	7.6	49	0.0	0
B02. Change	0.9	6	0.0	0
C. Property				
C01. Characteristic Property	4.3	28	0.0	0
C02. Psychology	01.	1	0.0	0
C03. Phenomenon Issue	1.2	8	0.0	0
D. Theory Method				
D01. Theory	10.0	64	10.9	7
D02. System	5.9	38	60.5	23
D03. Method	20.2	134	48.5	65
D04. Technique Strategy	6.4	41	29.2	12
E. Format Framework				
E01. Form Type Style Genre	0.1	1	0.0	0
E02. Model Criteria	16.4	104	45.1	47
E03. Languages	0.0	0	0.0	0
E04. Space	0.0	0	0.0	0
X. General Common				
X01. Place Name	0.0	0	0.0	0
X02. Period Time	0.0	0	0.0	0
X03. Relationship Interaction	1.4	9	0.0	0
X04. Result	1.2	8	0.0	0
X05. Aim	0.0	0	0.0	0
X06. Cause	0.1	1	0.0	0
X07. Effect	0.0	4	0.0	0
Total	100.0	638	35.2	225

Note. "% (PN)" shows the percentage of cite terms classified into proper nouns in each category and "n (PN)" shows the frequency of cite terms classified into proper nouns.

The second point relates to the percentages for proper nouns (column 4 of Table 1). The percentage of proper nouns is 35.3%. The category with the highest percentage of proper nouns is *A. Object* (50.0%). The category with the second highest percentage is *E. Format Framework* (44.7%). The category with the third highest percentage is *D. Theory Method* (38.6%).

Examinations About the Main Categories of Cite Terms

In contrast to the cite terms classified as *D. Theory Method* and *E. Format Framework*, some cite terms classified under other categories are domain independent (e.g., "The detection of

spammers" in *B. Action Function*, "the urgency of work task performance" in *C. Property*, and "the result of research" in *X. General Common*). There is a diversity of the content represented by the cite terms in *A04. Artifacts*. To clarify the types of cited content represented by the cite terms, we analysed the instances in *A04. Artifacts* according to proper nouns and common nouns. In the example sentences below, the cite expressions are indicated in bold print.

The cite terms classified under *A04. Artifacts* and common nouns can be divided into whether it refers to particular content in the cited document or to the whole cited document itself. As shown in Table 1, the number of cite terms categorised as *A04. Artifacts* and common nouns is 71, and the number of cite terms categorised as *A04. Artifacts* and proper nouns is 69.

The number of cite terms referring to the cited document itself is 55 of the 71 cite terms classified as *A04. Artifacts* and common nouns. An example is:

> *First, scientific literature could benefit from our in-depth up-to-date review about CRM CSFs: past reviews (e.g. Croteau & Li, 2003; Mendoza et al., 2007) are still valuable contributions, obviously, but they could lack the more comprehensive point of view stemming from over 10 additional years of academic research.*

In this example, the cited content is <past reviews>. This cited content refers to the cited document itself.

The other 16 instances of cite terms indicate particular content in the cited document. An example is:

> *Inspired by the relational charts for the presentation of relative indicators (see Braun et al. 1985; Schubert and Braun 1986), we show the plot of the relative citation rate (RCR, see Schubert and Braun 1986) versus relative usage rate (RUR) in this study.*

In this example, the cited content is <the relational charts for the presentation of relative indicators>. This is a part of the content in the cited document.

The cite terms classified as *A04. Artifacts* and proper nouns can be divided into whether the cite term refers to the document itself or an individual item. We identified 29 terms referring to the cited document itself. An example is:

> *As a result, they improved the F-measure by about 0.05% compared to Li and Sun (2014).*

In this example, the cited content is <Li and Sun (2014)>. This cited content refers to the cited document.

The other 40 instances indicate individual items, for example:

> *We randomly extracted 2,010 non-duplicate metadata records from two different digital collections: The UNT Library Catalog (University of North Texas Libraries, n.d.), and the Portal to Texas History (University of North Texas Libraries, 2016).*

In this example, <The UNT Library Catalog> and <the Portal to Texas History> are the cited content. These refer to individual entities referred to in the cited documents, but not to the cited documents themselves.

Table 2. The result of the classification of cite clauses

Category	%	n
Eventuality	50.0	489
General stative	39.4	386
Abstract entity	10.5	103
Total	100.0	978

Analysis of Cite Clauses

Overall Tendencies of Cite Clauses

Table 2 shows the result of the classification of the cite clauses using Situation Entity. Situation Entity at the first level comprises three categories: *Eventuality* constitutes half of the cite clauses; about 40% of the cite clauses are *General stative*; and about 10% are *Abstract entity*.

These two points can be drawn from the result. The first point is that the percentage of the cite clauses that represent subjective content (*Abstract entity*) is small. This may be because of the nature of scientific papers: authors do not tend to write subjective content in scientific papers. The second point is that the percentage of the cite clauses which represent term-specific content (*Eventuality*) is higher than the percentage of the cite clauses which represent general content (*General stative*). This indicates that, whilst the number of cite clauses that represent term-specific content are larger than the number of cite clauses that represent general content, the difference is not so large.

Examinations About the Main Categories of Cite Clauses

To understand the nature of cite clauses in more depth, we analyse examples of cite clauses in each category below. In these examples, the whole clause represents a cite expression.

Clauses of *Eventuality* are divided into two types. The first type is an instance of research, for example:

> *Zamani and Croft (2016, 2017) have also studied the impact of neural embeddings on language models in two consecutive work.*

In this example, the cited content is the act of the research but not the result of the research.

Some cite clauses introduce multiple instances of research into the citing document as cited content, for example:

> *The existing studies have typically approached the question using one of the following two methods: (1) by studying the unit of measurement itself, for instance in the case of citations by studying the motivations of scientists for choosing to reference or to not reference particular papers.*

This clause describes more than one study or act of research. This instance shows that the act of research can be cited content. The content conveyed by the citation includes the research activity itself, and the research activity can be an object of discussion in the citing document.

The second type is an instance of domain independent events, for example:

> *For instance, the European Research Council (ERC) supported Emerging Research Areas and their Coverage by ERC-Supported Projects (ERACEP) to identify emerging*

areas by topic and to analyze the extent to which ERC-funded projects contributed to these emerging areas in 2009.

The content described in this clause is an act in daily life but not an act of research.

Two types of *General stative* in cite clauses can be discerned. The first type is a general phenomenon:

It is impossible to estimate the magnitude of the trade-off.

ADR is the fourth leading cause of death in the United States.

These clauses describe general phenomena about the subject of research ("to estimate the magnitude of the trade-off" and "ADR").

The second type is the meaning of a term, for example:

Doc2Vec is an unsupervised learning of continuous representations for variable-length pieces of texts, such as sentences, paragraphs or entire documents.

Co-citation is defined as the frequency with which two units are cited together.

These clauses describe the nature or definition of the subject in research ("Doc2Vec" and "Co-citation").

Abstract entity, which comprises about 10% of all the cite clauses in our data, can be divided into three subtypes—judgement, possibility, and necessity. Here are two examples for the first type of judgement:

Stages of tasks play a crucial role in influencing users' application of information-seeking strategies

… training method for each user is more efficient in terms of recall and precision than the training method for every user

The predicates ("play a crucial role" and "is more efficient") in these cite clauses represent the subjective judgement of writers. Examples for the second category of possibility are as follows:

Any ambiguity at the morphological level will have an impact on the syntactic level.

A larger author team might reduce the need for rewriting of a paper as specialization ensures all parts of the paper are completed.

For the type of possibility, the content is expressed as a subjective content by the word such as "will" and "might" in the example sentences above. These words indicate the possibility of the phenomena described by the clauses. Examples for the last type, necessity, are as follows:

Therefore, both the syntactic information and semantic information must be used when comparing two sentences.

In addition, managers should strive to assess BD initiatives, and their usefulness, by setting appropriate metrics.

These two examples indicate necessity by the predicates with the words "must" and "should".

CONCLUSIONS

In this research, we have identified terms and clauses as the main types of linguistic expressions representing cited content and presented the distributions of cite units for two aspects: linguistic expressions and content. Regarding the aspect of linguistic expressions, we

pointed out that 638 of the cite units were expressed at the term level (39.5%) and 976 were expressed at the clause level (60.4%). Regarding the second aspect of content, we classified the cited content according to the type of linguistic expressions.

The analysis of the content of cite terms found that more than half of the cite terms refer to domain specific content. The analysis of the main types of cite terms revealed the following:

1. We identified two types of content in cite terms which were common nouns: particular content expressed or referred to in the cited document, and the cited document as a whole.
2. We identified two types of the content in cite terms which were proper nouns: particular content in the cited document, and the cited document as a whole.

Analysis of the cite clauses found the following:

1. Cite clauses of subjective content were in the minority, and cite clauses of objective content were in the majority of cite clauses.
2. The percentage of cite clauses representing objective and term-specific content was close to the percentage of cite clauses representing objective and general content (difference of 10%).

These results suggest a tendency that researchers generally cited both of past events and abstract ideas and not cited subjective content. Analysis of the main types of cite clauses identified seven categories: instance of research, instance of domain independent events, general phenomenon, meanings of a term, judgement, possibility and necessity.

The main contribution of this research was the description of the cite units from the perspective of linguistic expression and cited content. The data is limited to journal articles in library and information science. We also focused on sentences with a citation anchor for this initial study of cited content. Although we used sentences as a basic unit of linguistic expressions that corresponds to the content unit here, different units of linguistic expressions can represent coherent content units, especially the units of linguistic expressions are larger than sentences. We are planning to carry out a task of identifying the larger unit of linguistic expressions than sentences that represent content unit, by using methods of discourse structure analysis such as the one for identifying Elementary Discourse Unit.

REFERENCES

Abdallah, Y., Zhendong, N., Tarus, J. K., & Arshad, A. (2019). A survey on sentiment analysis of scientific citations. *The Artificial Intelligence Review, 52*(3), 1805–1838.

Alomari, J. S. (2016). Ontology for academic program accreditation. *International Journal of Advanced Computer Science and Applications, 7*(7), 123–117.

Anderssone, A., & Gronlund, A. (2009). A conceptual framework for e-learning in developing countries. *The Electronic Journal of Information Systems in Developing Countries, 38*(1), 1–16.

Asher, N., Benamara, F., & Mathieu, Y. Y. (2008). Categorizing opinion in discourse. In *Proceedings of the 2008 Conference on ECAI 2008: 18th European Conference on Artificial Intelligence* (pp. 835–836).

Bonab, H., Zamani, H., Learned-Miller, E. G., & Allan, J. (2018). Citation worthiness of sentences in scientific reports. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1061–1064).

Bornmann, L., Wray, K. B., & Haunschild, R. (2019). Citation concept analysis (CCA): A new form of citation analysis revealing the usefulness of concepts for other researchers illustrated by two exemplary case studies including classic books by Thomas S. Kuhn and Karl R. Popper. *Scientometrics, 122*(2), 1051–74.

Cohan, A., Ammar, W., van Zuylen, M., & Cady, F. (2019). Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3586–3596).

Dik, S. C. (1997). *The theory of functional grammar: The structure of the clause* (2nd ed.). M. de Gruyter.

Douglas, L. (1998). Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM, 38*, 32–38.

Elizarova, A., Khaydarov, S., & Lipachev, E. (2017). Scientific documents ontologies for semantic representation of digital libraries. In *Second Russia and Pacific Conference on Computer Technology and Applications* (pp. 1–5).

Färber, M., & Sampath, A. (2008). Determining how citations are used in citation contexts. In A. Doucet, A. Isaac, K. Golub, T. Aalberg, & A. Jatowt (Eds.), *Digital libraries for open knowledge* (pp. 380–383). Springer.

Friedrich, A., Palmer, A. (2014) Situation entity annotation. In *Proceedings of LAW VIII: The 8th Linguistic Annotation Workshop* (pp. 149–158).

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., & Schneider, L. (2002). Sweetening ontologies with dolce. In *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web* (pp. 166–181).

Giancarlo, G., & Gerd, W. (2010). Using the unified foundational ontology (UFO) as a foundation for general conceptual modeling languages. In *Theory and Applications of Ontology: Computer Applications* (pp. 175–196).

Grenon, P., Smith, B., & Goldberg, L. J. (2004). Biodynamic ontology: Applying BFO in the biomedical domain. *Studies in Health Technology and Informatics, 102*, 20–38.

Halliday, M. A. K. (1994). *An introduction to functional grammar* (2nd ed.). Edward Arnold.

Hernádez-Alvarez, M., & Gomez, J. M. (2016). Survey about citation context analysis: Tasks, techniques, and resources. *Natural Language Engineering, 22*(3), 327–349.

Herre, H., Heller, B., Burek, P., Hoehndorf, R., Loebe, F., & Michalek, H. (2001). *General formal ontology (GFO): A foundational ontology integrating objects and processes*. University of Leipzig. https://www.onto-med.de/sites/www.onto-med.de/files/files/uploads/Publications/2007/gfo-part1-v1-0-1.pdf

Holmes, F. L. (1987). Scientific writing and scientific discovery. *Isis., 78*, 220–235.

Ihsan, I., & Qadir, M. A. (2019). CCRO: Citation's context & reasons ontology. *IEEE Access, 7*, 30423–30436.

Jurgens, D., Kumar, S., Hoover, R., McFarland, D. J., & Jurafsky, D. (2016). Citation classification for behavioral analysis of a scientific field. *CoRR*, abs/1609.00435.

Kalemi, E., & Martiri, E. (2011). FOAF-academic ontology: A vocabulary for the academic community. In *2011 Third International Conference on Intelligent Networking and Collaborative Systems* (pp. 440–445).

Kang, I.-S., & Kim, B.-K. (2012). Characteristics of citation scopes: A preliminary study to detect citing sentences. In *Computer Applications for Database, Education, and Ubiquitous Computing* (pp. 80–85). Springer.

Kitamura, Y., Sano, T., Namba, K., & Mizoguchi, R. (2002). A functional concept ontology and its application to automatic identification of functional structures. *Advanced Engineering Informatics, 16*(2), 145–163.

Ko, Y. M., Song, M. S., & Lee, S. J. (2016). Construction of the structural definition-based terminology ontology system and semantic search evaluation. *Library Hi Tech, 34*(4), 705–732.

Leydesdorff, L. (1987). Towards a theory of citation? *Scientometrics, 12*(5–6), 305–309.

Li, X., He, Y., Meyers, A., & Grishman, R. (2012). Towards fine-grained citation function classification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2013* (pp. 402–407).

Lin, C.-S. (2018). An analysis of citation functions in the humanities and social sciences research from the perspective of problematic citation analysis assumptions. *Scientometrics, 116*(2), 797–813.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science, 5*(1), 86–92.

Niles, I., & Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems* (pp. 2–9).

Palmer, A., Ponvert, E., Baldridge, J., Smith, C. (2007) A sequencing model for situation entity classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (pp. 896–903).

Pertsas, V., & Constantopoulos, P. (2016). Scholarly ontology: Modelling scholarly practices. *International Journal on Digital Libraries, 18*, 173–190.

Reits, J. M. (2004). *Dictionary for library and information science*. Libraries Unlimited.

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.

Shum, S. B., Motta, E., & Domingue, J. (2000). Scholonto: An ontology-based digital library server for research documents and discourse. *International Journal on Digital Libraries, 3*(3), 237–248.

Small, H. G. (1978). Cited documents as concept symbols. *Social Studies of Science, 8*(3), 327–340.

Small, H. G. (2011). Interpreting maps of science using citation context sentiments: A preliminary investigation. *Scientometrics, 87*, 373–388.

Smith, C. S. (2003). *Modes of discourse: The local structure of texts*. Cambridge University Press.

Taşkin, Z., & Al, U. (2018). A content-based citation analysis study based on text categorization. *Scientometrics, 114*(1), 335–357.

Terziev, I., Kiryakov, A., & Manov, D. (2005). *D1.8.1 base upper-level ontology (BULO) guidance*. Ontotext Lab. https://www.sti-innsbruck.at/sites/default/files/sekt-d-1-8-1-Base_upper- level_ontology__BULO__Guidance.pdf.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). An annotation scheme for citation function. In *Proceedings of the 7th SIGDIAL Workshop on Discourse and Dialogue* (pp. 80–87).

Valenzuela, M., Ha, V., & Etzioni, O. (2015). Identifying meaningful citations. In *AAAI Workshop: Scholarly Big Data* (pp. 21–26).

Wang, M., Zhang, J., Jiao, S., Zhang, X., Zhu, N., & Chen, G. (2020) Important citation identification by exploiting the syntactic and contextual information of citations, *Scientometrics, 125*, 3109–3137.

Widyantoro, D. H., Khodra, M. L., Trilaksono, B. R., & Aziz, E. A. (2013). A multiclass-based classification strategy for rhetorical sentence categorization from scientific papers. *Journal of ICT Research and Applications, 7*, 235–239.