Manuscript of: Cheng, W.-N., & Khoo, C.S.G. (2021). Information structures in sociology research papers: Modeling cause–effect and comparison relations in research objective and result statements. *Journal of the Association for Information Science & Technology, 72*(11), 1367-1385.

Information Structures in Sociology Research Papers: Modeling Cause-effect and Comparison Relations in Research Objective and Result Statements[1]

Wei-Ning Cheng, and Christopher S.G. Khoo

Department of Library, Information and Archives, School of Management, Shanghai University, China;

Wee Kim Wee School of Communication & Information, Nanyang Technological University, Singapore

Author Note

Wei-Ning Cheng https://orcid.org/0000-0001-6197-1503

Christopher S.G. Khoo https://orcid.org/0000-0002-8072-1072

We have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to Christopher S.G. Khoo, Wee Kim Wee School of Communication & Information, 31 Nanyang Link, Singapore 637718. Email: chriskhoo@pmail.ntu.edu.sg

[1] This study is part of the Ph.D. research project of the first author.

Abstract

When writing a research paper, the author has to select information to include in the paper to support various arguments. The information has to be organized and synthesized into a coherent whole through relationships and information structures. There is hardly any research on the information structure of research papers, and how information structure supports rhetorical and argument structures. Thus, this study is focused on information organization in the Abstract and Introduction sections of sociology research papers, analyzing the information structure of research objective, question, hypothesis and result statements. The study is limited to research papers reporting research that investigated cause-effect relations between two concepts. Two semantic frames were developed to specify the types of information associated with cause-effect and comparison relations, and used as coding schemes to annotate the text for different information types. Six link patterns between the two frames were identified—showing how comparisons are used to support the claim that the cause-effect relation is valid. This study demonstrated how semantic frames can be incorporated in discourse analysis to identify deep structures underlying the argument structure. The results carry implications for the knowledge representation of academic research in knowledge graphs, for semantic relation extraction, and teaching of academic writing.

Keywords: knowledge graph, knowledge representation, social science research, semantic relations, discourse analysis, information structure, academic writing

Information Structures in Sociology Research Papers: Modeling Cause-effect and Comparison Relations in Research Objective and Result Statements

When writing a research paper, the author has to select the information to include in each section of the paper, and organize the information for ease of reading and comprehension. More than that, pieces of information are used to support various arguments that the author makes in the paper. The arguments and information have to be presented in text in a sequential or linear form, in a way that helps the reader to understand the arguments and to be persuaded of their validity. The sequence of persuasive or rhetorical functions represents the rhetorical structure of the text (Lim, 2011). Analysis of the structure of research papers is done mainly by researchers in applied linguistics. In particular, researchers in the subfield of genre studies analyze the rhetorical structure of research papers, often following Swale's (1990) *Creating a Research Space* (CARS) model. There are very few studies of the information structure of research papers and of information use in research papers. Studies of information use have concentrated on analysis of citations types and citation functions (e.g., Burbules, 2015; Lin, 2018; Stremersch, Camacho, Vanneste, & Verniers, 2015).

This study is part of a larger project to carry out discourse analyses of research papers in multiple academic disciplines, and to investigate relations between *rhetorical structure, argument structure* and *information structure*. *Rhetorical structure* represents the surface presentation structure of the text for the purpose of persuasion. However, the rhetorical purpose must be to persuade or convince the reader about *something*—that is, information and arguments. Our theoretical assumption is that *rhetorical structure* is built on the *argument structure*, which is itself built on the *information structure*. This study sought to identify relations between information structure and argument structure, to find out how information structure is used to explain or clarify an argument. A separate paper will explore the relationship between argument structure and rhetorical structure. We carried out information structure analysis of the Abstract and Introduction sections of sociology research articles from a knowledge representation perspective—identifying `concept->relation->concept` triples represented in the text and linking them into bigger semantic structures that can be considered knowledge graphs.

The Abstract provides an overview of the information content and argument structure of the research paper. The Introduction section also conceptualizes the overall research, usually with less information on the research results. Ahlstrom (2017) characterized it as a microcosm of the whole research paper. Flowerdew (1999) indicated that the Introduction section is one of the most difficult sections to write in a research paper because the author needs to convince the readers "of the importance of their research and the arguments they are putting forward." (p. 258).

Information structures in the Introduction sections of research papers are usually rather complex. In a preliminary study, we found it too tedious and time-consuming to code the whole Introduction

section as many overlapping information structures may be involved. Rather than be swamped with many details and lose sight of the big picture, we decided to focus on the most important statements in the Introduction section—the Research objective and Research result statements. *Sociology* was selected as the social science discipline to analyze as we expected that the information and discourse patterns found are more likely to be prevalent also in other social science disciplines, compared to disciplines such as linguistics, law and geography.

In a preliminary study, we found that different types of research are associated with different information structures. We identified five types of research studies: *Investigative research, Development and evaluation research*, *Historical analysis*, *Descriptive research*, and *Identification research* (Cheng, 2020). The types of research reflect different epistemic paradigms involving different types of concepts, issues and entities studied, and different types of knowledge sought, with implications for the research methods used. The research type of a research paper can be determined based on the research objectives and research method used (Cheng, 2020).

This study focused on sociology research articles reporting *Investigative research*, which we define as research investigating a *research relation* (usually a *cause-effect relation*) between two concepts or entities. Such studies typically employ quantitative research methods, but may employ a qualitative method. They can be characterized as adopting a positivist paradigm of research, rather than a postmodernist, constructivist, interpretivist or critical theory paradigm (Lincoln, Lynham, & Guba, 2011; Schwandt, 1998).

Different types of research are associated with different information structures as both the target concepts and the types of knowledge sought are different, implying different conceptual relations and different associated concepts. We represent information structure patterns using a frame-based representation based on Fillmore, Johnson, and Petruck's (2003) *semantic frames*, which they defined as "schematic representations of the conceptual structures and patterns of beliefs, practices, institutions, images, etc. that provide a foundation for meaningful interaction in a given speech community" (p. 235). Our semantic frames list the types of information that are relevant or expected when describing a particular situation (e.g., when describing a causal relation between two concepts), and the role each type of information plays in the situation.

From the preliminary study, we developed six research semantic frames that reflect common information structure patterns found in the initial sample of research papers: *Research-relation frame, Development and Evaluation frame, Descriptive frame, Comparison frame, Theory/model/framework frame*, and *Measurement frame*. This paper focuses on two important semantic frames found in papers reporting *Investigative research*: Research-relation frame and Comparison frame.

As an *Investigative research* study investigates a research relation between two concepts, the Research-relation frame must thus be instantiated at least once in the research objective of the study. The Research-relation frame is used as an analytic tool to identify:

- What type of research relation was investigated (i.e. whether *cause-effect, association, prediction*, etc.)?
- Which two concepts are linked by the research relation?
- Which of the other relevant types of information/roles are specified?

Comparisons are prevalent in investigative research studies. A cause-effect relation is often established from the result of a comparison. For example, two potential causal factors may be compared in terms of their effect and effect size, based on some criterion measure. Thus, this study made use of the Comparison frame to analyze the types of comparisons made in the study, which of the relevant types of information are specified, and how the comparison is related to the research relation—that is, how the instantiated Comparison frame is linked to the instantiated Research-relation frame.

The objectives of the study were:

1. to identify the characteristic information profiles of research objective statements (and related statements of hypothesis and research question) and research result statements found in the Abstract and Introduction sections of sociology research papers.
2. to identify link patterns between Comparison and Research-relation frames, showing the different ways in which comparisons are used to support the argument claim that the Research-relation is valid.

Example texts illustrate how the link patterns are realized in the text.

A methodological contribution of this paper to discourse analysis is the method of analyzing information structure of research papers using semantic frames. A repertory of six research semantic frames were developed, although this paper focuses on two frames that are important to investigative research. The results reported in this paper show the utility of this method of analyzing information structure.

The main application of the results of this study is in the knowledge representation of academic research in a machine-readable form. Academic research is currently reported in textual form in research papers, which can be understood and processed only by human readers. For example, synthesizing research results from multiple research papers are accomplished by human authors in literature reviews. There is a research thread in the area of digital libraries and text mining that seeks to represent research processes and results in a machine-processable knowledge representation such as knowledge graph (Gutierrez & Sequeda, 2019; Ehrlinger & Wöß, 2016)—to support semantic information retrieval, customized information extraction, generation of literature overviews, reproduction of research results, and updating of systematic reviews (Brack, Hoppe, Stocker, Auer, & Ewerth, 2020; Slaughter, Berntsen,

Brandt, & Mavergames, 2015). This emerging research area can be characterized as *scientific knowledge graph*. A major initiative is the Open Research Knowledge Graph project at TIB-Leibniz Information Center Technology and Natural Sciences and University Library, Germany.[2] Project researchers, Jaradeh, Oelen, Farfar, Prinz, D'Souza, Kismihók, Stocker, and Auer (2019), argue that "there is an urgent need for a more flexible, fine-grained, context sensitive and machine actionable representation of scholarly knowledge and corresponding infrastructure for knowledge curation, publishing and processing." The project noted that "the structured description of research contributions is no easy task. ... You need to decide at what level of granularity you want to describe a research contribution, the addressed problem, its results and employed material and methods. ... To address this issue, ORKG supports the possibility of creating templates that specify the structure of content types, and using templates when describing research contributions."[3] This paper contributes to the issue of designing knowledge representation templates for research objectives and results. Current research efforts are focused on representing scientific, particularly biomedical, research knowledge. Inevitably, research efforts will expand to social science research representation, which our study contributes to. Though our analysis is of sociology research papers, we expect the information structures identified to be found in other social science as well as scientific disciplines, as cause-effect and comparison relations are prevalent in all research disciplines.

Our study carries implications for information extraction from research papers, especially semantic relation extraction, to generate scientific knowledge graphs. Previous studies have focused on extracting binary relations between two entities, for example the SemEval-2018 Semantic Relation Extraction task (Gábor, Buscaldi, Schumann, QasemiZadeh, Zargayouna, & Charnois, 2018). Zhou, Zhong, and He (2014) called for more studies in the extraction of complex relation structures involving more than two entities/arguments. Our study indicates what relation structures can usefully be extracted to represent research objectives and results, as well as information structures that link argument claims and supports. The information patterns derived in our study can also be taught as research, writing and thinking patterns to undergraduate and graduate students, as part of academic skills instruction.

Theoretical Background and Related Work

Text structure is complex and multi-layered. Text analysis carried out in the context of genre, social context and communicative purpose is referred to as *discourse analysis*, sometimes defined as the analysis of *language in use* (Gee 2014; Johnstone, 2017; Miles, 2010). *Discourse analysis* is a fuzzy concept, and researchers in different fields have defined it variously, and developed different discourse

[2] https://projects.tib.eu/orkg/
[3] https://projects.tib.eu/orkg/documentation/

analysis methods to address different research questions. Schiffrin, Tannen and Hamilton (2001) categorized the extant definitions of *discourse analysis* into: 1) anything beyond the sentence; 2) language use; 3) a broader range of social practice.

One type of discourse analysis is rhetorical structure analysis. Researchers in genre studies have carried out rhetorical structure analyses of research papers in many academic disciplines using a version of Swale's (1990) CARS model—a framework of rhetorical moves and more specific steps (which can be characterized as rhetorical functions). Research objective, hypothesis, research question, and research result are analyzed as rhetorical functions in the CARS model. However, they do imply certain arguments. For example, the research objective statement carries the implicit claim that it is well-founded (i.e. derived from theory, the literature, or observations) and worth investigating (because it addresses a research gap, is a novel idea of interest to a community, or is an important issue). When viewed as an argument claim, the research objective has to be supported with supporting arguments, such as research gap and importance of the topic (topic centrality), as well as information content.

The need to explain the rhetorical and argument structure of a text at least partly by the information the author seeks to communicate has been indirectly acknowledged by linguists and researchers in natural language processing. For example, Kwan, Chan and Lam (2012) analyzed the rhetorical structure of the literature review sections of information systems journal articles, comparing articles favoring a human behavior and organizational focus, and articles favoring a technical focus (i.e. algorithms, mathematical models and system design). They found that literature reviews in the two research paradigms tend to present different types of information even for the same rhetorical function. Human behavior papers tend to evaluate theories, concepts and variables, whereas technical papers evaluate algorithms, models and techniques. They identified "making inferences" as an important strategy for evaluating previous work, often used to support the rhetorical function of *formulating a research hypothesis*. This evaluation strategy is supported by two kinds of information structure: how a variable or concept may cause another, and speculation about the relation between two concepts. These two information structures are represented as the Research-relation frame in our study. However, Kwan et al. analyzed the information types and information structure qualitatively and informally, as a supplement to rhetorical structure analysis. Our analysis sought to identify micro-level generic semantic structures that support an argument, as an intermediate step towards the surface rhetorical structure.

The main contribution of this study to *discourse analysis* is the method of information structure analysis drawn from the field of knowledge representation. Researchers in the field of applied linguistics do make use of the concept of *information structure* (sometimes called *information organization* or *information management*). However, their concept of information structure refers to "information distribution" within a clause or sentence, and is closely tied to the sentence syntactic structure (Lovejoy,

1991, Roberts, 2012). In the analysis of information distribution, information content in a clause or sentence is analyzed into two parts: the given or old information (usually appearing in the first part of the sentence), and the new information (in the rest of the sentence). This *given-new* relation has also been referred to as *theme-rheme*, *topic-comment*, and *focus/(back)ground* (Lovejoy, 1991; Roberts, 2012).

In our discourse analysis, related pieces of information can be distributed across multiple sentences. Furthermore, we do not link the information structure to the sentence syntactic structure, but to the communicative purpose, that is, the main argument claim (e.g., research objective and research result). The pieces of information in each section of a research paper, and the network of relations between them comprise the information structure of the section. An information structure represents the main types of research information reported in the paper, and how they are linked into meaning structures (referred to as semantic frames in this study). The semantic frames (representing higher levels of meaning than individual concepts) are themselves connected to one another to support the argument structure.

To model information structure, we adopted Frame Semantic Theory (Fillmore, Johnson, & Petruck, 2003; Fillmore, 1968) which models the types (roles) of information relevant to a central concept (often expressed as "verbs"). Frame Semantic Theory can be considered a knowledge representation approach based on the assumption that some words (especially verbs) evoke a frame of background knowledge relevant to the particular concept. In our broader project, we focus on certain important concepts (prevalent in research papers) and their associated frame of background knowledge: Research-relation, Development and evaluation, Description, Comparison, Theory/model/framework and Measurement. This paper, however, focuses just on the Comparison and Research-relation frames.

The information structure intertwines with the argument structure to support coherence. Lovejoy (1991) noted that information structure is related to cohesion and coherence of the text. Cohesion refers to ties between parts of a sentence and between sentences. Cohesive devices or markers include references, conjunction, ellipsis, and lexical ties (e.g., repetition and synonyms) (Halliday & Hasan, 1976). Coherence includes cohesive ties but also conceptual relations between information entities. From a cognitive perspective, *coherence* can be characterized as "a quality of the mental representation of texts that is created by the reader" (McNamara, Crossley, & McCarthy, 2010, p. 60). That is, the reader links together pieces of information expressed in the text into a mental structure that makes sense to the reader. Thus, we investigated semantic relations between pieces of research information expressed in the text—that the reader is expected to infer when reading the research paper.

There is a related body of research arising from Rhetorical Structure Theory (RST) developed by Mann and Thompson (1988). It is a theory of text organization that provides an explanation of coherence by deriving a linked tree structure of a text. RST analysis divides text units into two parts: the nucleus indicating the primary part of the text, and the satellite indicating the secondary part. Rhetorical relations

(also referred to as coherence relations) are used to indicate different relations between nucleus and satellite. In short, in RST, the rhetorical structure (derived from the sequential text presentation) is combined with semantic relations to explain coherence of the text. As the RST analysis is carried out from the beginning to the end of the text, one can argue that the focus of RST is on the rhetorical structure, with some semantic and argument relations added to strengthen the structure. In contrast, the current study has focused on information structure separate from the presentation order.

A recent version of RST defines 21 relations (Mann & Taboada, 2021). Six of the relations are included in our semantic frames. For example, *evidence* in RST indicates a claim and relevant information for persuading the reader's trust in the claim; *condition* shows a relation between a situation and a particular condition (called *qualifier* in our Research-relation frame). RST is developed for analyzing any kind of text and therefore defines generic relations that are broadly applicable. In contrast, we define more specific types of relations that characterize research arguments.

Framework: Research Information Model

Research-relation Frame

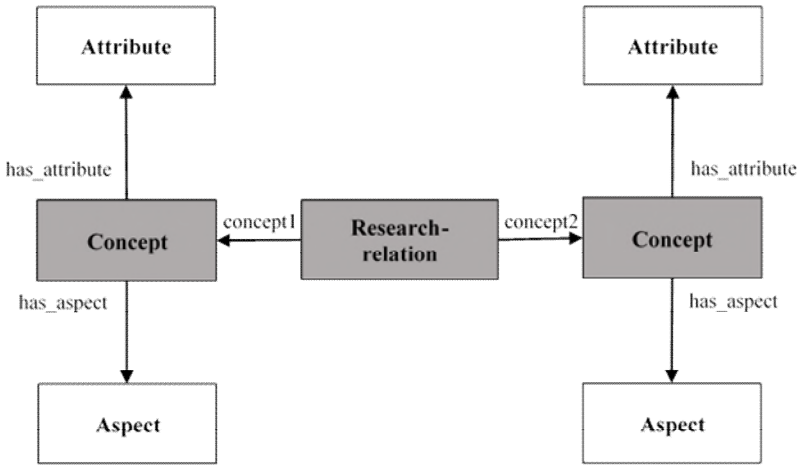
The cause-effect relation is arguably the most important relation in research, as many research studies seek causal knowledge by identifying the *cause* of a particular phenomenon, or the *effect* of particular factors. However, causality is difficult to establish as, ideally, it requires a randomized controlled experiment. Therefore, some studies focus on identifying different types of associations (e.g., correlation, co-occurrence, prediction, or a vague association). We grouped all these as *research relations*. Of course, not all research studies are focused on investigating research relations. This study has focused on research papers that seek to identify research relations (mainly cause-effect relations), as reflected in their research objective statements.

Papers reporting *Investigative research* usually provide additional information that clarify or support the research relation. We represent the various types of information in a Research-relation semantic frame (see Figure 2). A few of these information types (i.e. evidence, context, modality and polarity) are adapted from Ou, Khoo and Goh's (2007) Variable-based Framework. The rest are developed based on an initial sample of 20 research papers (not part of the 50 papers analyzed for this paper). The information types in the Research-relation frame are represented as elements in a metadata scheme, and used as tags for annotating text. The metadata elements map directly to relations or roles in the ontology diagram of Figure 2. Note that the Research-relation frame may be linked to other semantic frames. In particular, concept instances in the Research-relation frame may be linked by relations in the Comparison frame, described later.

The Research-relation is rather complicated with many roles. We describe the important parts of the frame here. The core of the frame is the *Research-relation* concept (or class) that is related to two concepts—the *cause* concept and the *effect* concept. To make the frame more general and applicable also to associative relations, we label the relations *concept1* (i.e. cause) and *concept2* (i.e. effect), as illustrated in Figure 1.

Figure 1

The Research-relation Concept Linked by Relations concept1 and concept2 to Two Relevant Concepts



The Research-relation concept has subclasses: *Cause-effect, Correlation, Prediction, Co-occurrence,* and *Association.* If the Research-relation is specialized to Cause-effect, then the relation *concept1* (in Figure 1) can be specialized to *cause,* and *concept2* to *effect*. As illustration, consider the following research objective:

how social media *affect* traveler behavior on hotel websites

This can be represented as:

[social media] <-(cause)– [**Cause-effect.*affect***] –(effect)-> [traveler behavior]

The distinction between Cause-effect, Association and Predictive relation is fuzzy. We observed that most of the Research-relations are casually expressed as causal relations (i.e. without rigorous experimental tests or philosophical justifications). Sometimes an Association relation is expressed for linguistic variety, when it is clear the author meant a causal relation.

Each concept linked to the *Research-relation* may have additional *attributes* (internal features) and *aspects* (external features) specified. Extending the research objective example above, the author may

compare the effect of hotel websites *with* embedded social media to those *without* embedded social media. This can be represented as:

[hotel websites] –(has_attribute)-> [embedded social media]

[hotel websites] –(has_attribute)-> [no embedded social media]

We observed from our coding experience that the distinction between *attribute* and *aspect* is fuzzy: an *attribute* is an intrinsic property of the Concept instance (entity), whereas an *aspect* reflects the interest or perspective of the researcher. However, an aspect may sometimes be perceived as intrinsic to the entity.

A *subclass* tag suggests that multiple subtypes of a broader Concept are investigated in separate cause-effect relations to determine which subclass is associated with which value of the effect concept:

My concern is with how concept1*social mobility — both at* subclass1*the individual level and the country level — affects* concept2*class identification.*

Instance refers to the words in the text that indicate a Research-relation. An instance can also indicate a more specific type of research relation or comparison relation. If the instance words indicate a positive or negative direction to the Research-relation, then they are coded as *Polarity.negative.*

The *size* (magnitude) of the relation is usually expressed qualitatively, which is often an instance of a Comparison frame:

Findings reveal that there are size*strong differences in normative climates across countries.*

The most important type of *modality* is negation, indicating that the relation does not hold. However, *modality* can also reflect degrees of certainty and definiteness, and other attitudes towards the research relation. It is often represented by modal verbs[4], but can be indicated with modal nouns (e.g., *possibility*), adjectives (e.g., *probable, unlikely*), and adverbs (e.g., *consistently, generally*).

Explanation refers to the underlying explanation or hypothesized theoretical mechanism that makes the Research-relation plausible. We found that the author may also point out the implications (including theoretical implications) of a Research-relation:

concept2*Daily spillover findings were largely unaffected by* concept1*parents ' neuroticism, suggesting that* explanation*parents' day-to-day fluctuations in negative mood, not average levels of negative affectivity, promoted spillover.*

The distinction between underlying explanation and theoretical implication is fuzzy, and thus we code both as *explanation*. An *explanation* also tends to provide more fine-grained details:

Moderator indicates a variable of less interest that modifies the effect of *Concept1* (cause concept):

[4] https://dictionary.cambridge.org/grammar/british-grammar/modal-verbs-and-modality; https://www.det.nsw.edu.au/eppcontent/glossary/app/resource/factsheet/4091.pdf

> *I examine the intersection of* moderator*gender and race in the effect of* concept1*marriage.*

The author may express an interaction between two variables:

> *Other controlling variables include the interaction between* moderator*race (in the pooled sample analysis only) and* moderator*potential experience.*

The author may also indicate controlled variables and confounding variables:

> *Associations persisted after controlling for* moderator*a range of work and family characteristics, and there was no evidence of mediation by* mediator*family socioeconomic status, maternal age, or job quality.*

Interaction, controlled and confounding variables are all coded as *moderator* variables, as they are related.

Another related element is *qualifier* which specifies a condition when the Research-relation will or will not hold. A *qualifier* is thus essentially a *moderator* variable, with a value of the variable held constant for the study:

> *Reciprocal influences were* qualifier*not confined to one period of parenting but continued as children grew older. Associations persisted after* moderator*controlling for a range of work and family characteristics ...*

In other words, a qualifier is a moderator variable that is not manipulated in the study, but its value is assumed fixed for the target population. The *moderator* and *qualifier* roles may be indicated as negated, suggesting that the roles can be divided into *positive* and *negative*.

A *mediator* or intervening variable is like a bridge between Concept1 and Concept2, and helps to explain the Research-relation between Concept1 and Concept2. If a *mediator* variable is specified, then the *mediator* has a direct relation with Concept2, whereas Concept1 has an indirect relation to Concept2.

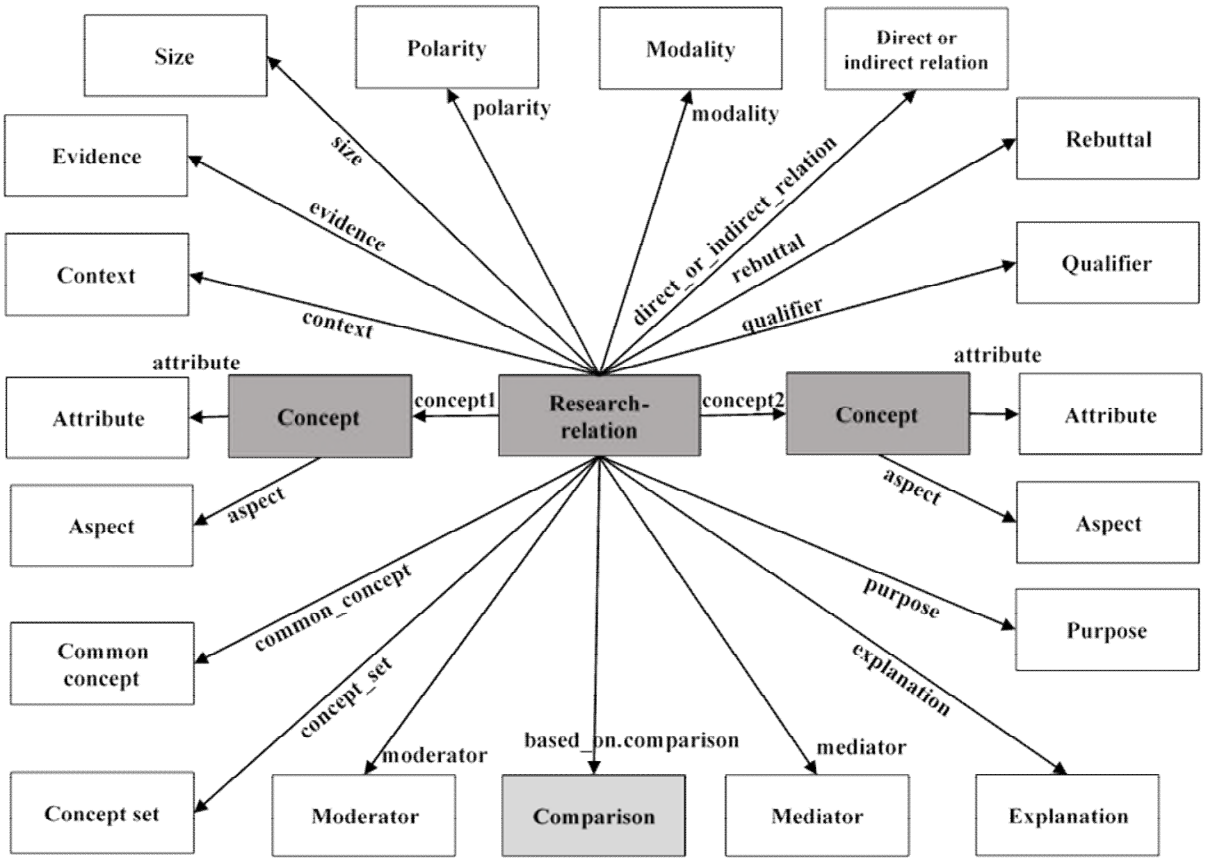
During the inter-coder analysis, we identified the following additional roles that can usefully be added to the Research-relation frame:

- *Common concept*—indicating that Concept1 and Concept2 are the same. The Research-relation is found to link different attributes/aspects of the Common concept.
- *Concept set*—indicating more than two concepts. Research-relations are investigated between every pair of concepts in the Concept set.

Purpose and *rebuttal* roles are hardly found in our corpus. They are more likely to be found in the literature review, theory/model/framework, and result sections of a research article.

Figure 2

The Research-relation Frame



Comparison Frame

Comparisons are often used to support a cause-effect relation, often by comparing two attributes of the *cause* concept (or two value categories of a *cause* attribute) and measuring an attribute value of the *effect* concept. One objective of this study is to identify the various configurations in which a Comparison frame is often linked to a Research-relation frame, which we refer to as a *link pattern*.

The types of information in the Comparison frame are shown in Figure 3. The comparison result may be based on measuring some criterion attribute (different from the attributes being compared). This is indicated by the *measure* role:

> *However,* $_{concept1}$*mothers had* $_{result.qualitative\_value}$*slightly less* $_{measure}$*pure free time than* $_{concept2}$*fathers and were more likely to* $_{difference1}$*combine leisure with unpaid work or spend time in leisure with children.*

We use *measure* to also indicate the measurement instrument or method used to measure the difference between *concept1* and *concept2* that are compared.

The *result* of the comparison is usually specified qualitatively (e.g., *same, similar*, and *concept1 is better*), and seldom quantitatively in the Abstract and Introduction section. Sometimes the comparison result is specified as an attribute value associated with *concept1* (indicated as *difference1*) or *concept2* (indicated as *difference2*). These features are illustrated in the following example:

> *… testing the claim that measure leisure with children and family is more likely to be experienced as difference1 leisure for concept1 men and difference2 work for concept2 women.*

The *comparison result* can also indicate what is *common* or the same between *concept1* and *concept2*:

> *Results showed that subclass1 mothers and subclass2 fathers spent common the same measure amount of time on leisure activities.*

So, *common, difference1* and *difference2* are actually sub-relations of *has_result*. Furthermore, the related concepts Common, Difference1, Difference2 are *part_of* the Result concept.

From the result/conclusion, the author may infer a Cause-effect relation (linked to a Research-relation frame), or a theory, model or framework (linked to a Theory/model/framework frame, not covered in this paper). These are indicated by *has_inference* elements.

To determine a research relation between quantitative (continuous) variables, the researcher may check for correlation or co-variation of the values of *Concept1* and *Concept2*, for example:

> *The longer concept1 the time interval between intention formation and the action is, and the greater spatial distance to a destination is, the higher probability to concept2 change behaviors.*

Co-variation can be viewed as a kind of comparison. However, in this study, we consider only comparisons between categories.

An author may also compare the results with expectation, hypothesis or previous results. Such comparisons are included in our analysis, but as they are usually found in research result statements, they rarely occur in Introduction sections.

A potential source of coding inconsistency is implied comparisons. A cause-effect relation always implies a comparison between different attribute values of the cause concept. For example, “Smoking causes cancer” implies a comparison between smoking and non-smoking. As we do not want to automatically annotate a comparison every time there is a cause-effect relation, the following would not be coded for comparisons:

> Smoking causes a higher probability of cancer.
>
> Smoking increases (raises) the probability of cancer.
>
> Smoking makes lung cancer more likely.
>
> A smoker is more likely to get cancer.

However, we coded a comparison if it is explicitly indicated or is emphasized, for example:

> The likelihood of cancer is higher for smokers compared to/versus/relative to non-smokers.

Clearly, this is a fuzzy area where coding inconsistencies are more likely.

Comparative adverbs and adjectives, for example *higher*, *more* and *relative* indicate comparisons, and we code them as such:

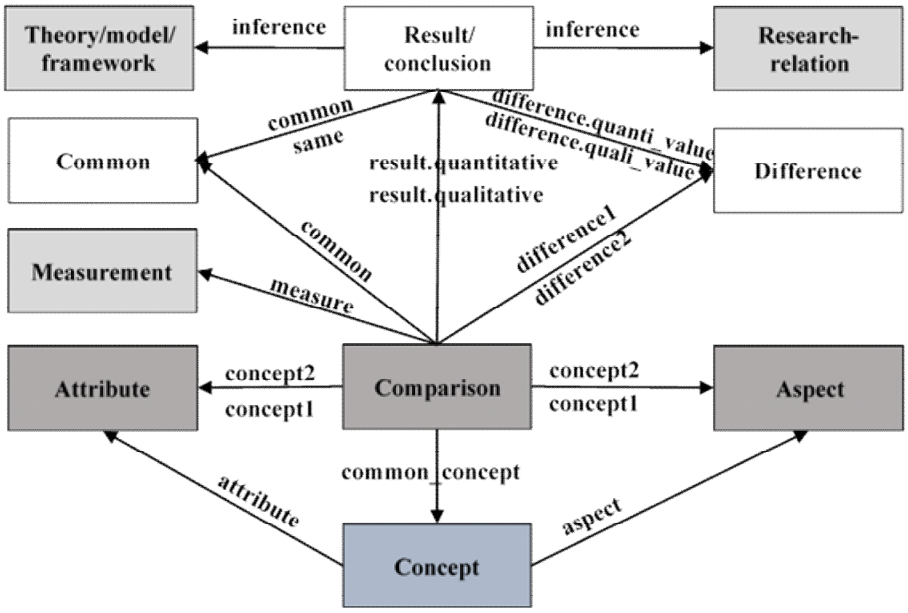
> … *a [Web] platform for creating unique co-creation of experiences, allowing tourists to become more physically and emotionally engaged in the planning of their vacations.*

The example statement implies a comparison to the case of no Web platform.

After inter-coder analysis, we decided to add a *common concept* role to refer to the concept whose attributes are compared (not to be confused with *common* which is a result of the comparison).

Figure 3

The Comparison Frame



Method

Corpus

We analyzed 50 sociology research articles reporting *Investigative research*, which sought to investigate cause-effect relations. Twenty of these articles were coded by both authors, and used for the inter-coder reliability analysis. Coding of the additional 30 articles was split between the authors. The 50 articles were sampled from 10 sociology journals with the highest impact factor in InCites Journal

Citation Reports: *American Journal of Sociology*, *Annals of Tourism Research*, *Cornell Hospitality Quarterly*, *European Sociological Review*, *Gender Society*, *Information Communication Society*, *Journal of Marriage and Family*, *Social Networks*, *Qualitative research*, and *American sociological review*. The articles were published in late 2015 or early 2016 volumes of the journals. Only articles reporting research that involved data analysis were included.

Coding scheme

The roles in the Research-relation frame (Figure 2) and Comparison frame (Figure 3) are represented as elements in a metadata scheme, and used as tags for annotating research paper texts. The papers in the corpus, originally in HTML or PDF format, were converted to XML format, and the annotated using XML tags using the oXygen XML editor software.[5] The text spans annotated are mainly noun phrases and clauses, but can be any phrase and even single words. The following resources that support the XML tagging and display are available from DR-NTU (Data) (the data repository of the Nanyang Technological University) doi:10.21979/N9/LD3EBQ:

- XML schema file to support the XML tagging and validation
- Cascading stylesheet file to display the annotated text in a Web browser
- Sample annotated text file
- documentation for the tag elements
- OWL/Turtle file that represents the semantic frames as classes and properties in an ontology, which can be instantiated with key phrases and concepts tagged in the research papers.

Inter-coder reliability

To estimate inter-coder reliability, a sample of 20 articles from the corpus were coded by both authors. The Jaccard similarity coefficient was used as the measure of inter-coder agreement. It measures the amount of overlap between two sets of elements, each set representing the elements identified by one coder.

In determining agreement, *concept1* is conflated with *attribute1, aspect1* and *subclass1*; similarly *concept2* is conflated with *attribute2, aspect2* and *subclass2*. This is because an attribute, aspect and subclass are, of course, also concepts. A complex noun phrase may be coded as a *concept* by one coder, or split more finely into *concept+attribute*, *concept+aspect* or even *concept+attribute+aspect*, depending on how fine-grained the coding is. If a text span is coded as an *attribute*, *aspect* or *subclass*, it implies that there is a related parent *concept* specified within the same sentence or in the previous sentence. *Instance* and *polarity* are also conflated, as many keywords that signal a *polarity* (e.g., *increase*, *decrease*) also indicate a Research-relation at the same time.

[5] https://www.oxygenxml.com/xml_editor.html

There is high agreement in coding the Research-relations together with *concept1* (cause concept) and *concept2* (effect concept). A Jaccard coefficient of 0.90 was obtained for the Cause-effect relation, 0.86 for the Association relation, and 0.80 for Research-relation *instance* (together with *polarity*). *Size* (0.83) and *context* (0.81) were also reliably coded. The rest of the information types have moderate and low agreements, reflecting some confusion between coders. The confusions between coders are generally as follows:

- Subclasses of Evidence and Context information overlap: *research method* (Evidence), *data source* (Evidence) and *target population* (Context); *time* (Evidence) and *temporal* (Context).
- *Qualifier* and *explanation*.
- *Modality*, *polarity* and *size*.

The error analysis has helped to clarify the meaning of some elements, and to carefully distinguish between potentially confusing elements. However, the core elements of the Research-relation frame (i.e. Cause-concept –(cause)-> Effect-concept) is reliably identified.

High agreement was obtained for identifying Comparison relations (Jaccard coefficient of 0.96). Moderate agreement was obtained for coding *instance* keywords, *common concept* and comparison *result*. As indicated earlier, some comparison relations are implicit, and the distinction between implicit and explicit comparisons are fuzzy.

Results

Percentage and Frequency of Research-relation Elements and Comparison Relations

The percentage of research papers (N=50) containing each Research-relation element and Comparison relation, and the average number per paper are listed in Table 1 (for Abstracts) and Table 2 (for Introduction sections). Only two Abstracts contain a research hypothesis statement, and one a research question statement, and so Table 1 reports the statistics only for research objective and research result statements. The percentages that are substantially higher for research objective or research result are indicated in bold print. Most of the research objective and research result statements in the Abstracts contain a Research-relation, mainly *cause-effect* relation. Comparing research result versus research objective statements, a research result statement is more likely to report an association relation. Not surprisingly, a research result statement is more likely to contain the following details: effect size, modality, moderator or qualifier variables, mediator, evidence and underlying explanation. However, the research objective is more likely to specify *context* information.

Looking at the statistics in Table 2 for Introduction sections, we find that only about 20% (N=11) of the Introduction sections contain a research result statement, 40% contain one or more hypotheses, and 30% contain one or more research questions. Most of the research objective, hypothesis, research

question and research result statements contain a Research-relation, again mainly *cause-effect* relation. About one-third of the research objective, research question and research result statements contain an *association* relation. The *association* relation is seldom found in hypothesis statements, which usually hypothesize the stronger *cause-effect* relation.

Examining the other elements, we find that *modality* is more likely to be found in the hypothesis (62%) and research result statements (55%), usually indicated by modal adverbs *may*, *significant*, etc. *Moderator/qualifier* variables are more often found in research result statements (45%). As with Abstracts, *context* information is more often found in research objective statements. Hypothesis statements are more likely to include underlying *explanations* (42%), because hypotheses are often derived from theory (theoretical explanations).

Looking at the statistics for Comparison relations in Tables 1 and 2, we find that comparisons are found in research result statements in 67% of Abstracts and 64% of Introduction sections, but only about 20% for research objective, hypothesis and research question statements.

Table 1

Percentage of Abstracts containing each Researc- relation element and Comparison relation, and average number per Abstract (N=50 Abstracts)

	Research objective (N=42 Abs)		**Research result (N=49 Abs)**	
Research-Relation Element	**% of Abs**	**Avg no. per Abs**	**% of Abs**	**Avg no. per Abs**
Research-relation (including concept1 and concept2)	88% (37 of 42)	2.38	92% (45 of 49)	**4.51**
• **cause-effect**	64% (27)	1.64	69% (34)	**2.73**
• **association**	26% (12)	0.74	**41% (20)**	1.45
• **prediction**	0	0.00	10% (5)	0.29
• **correlation**	0	0.00	2% (1)	0.04
instance (including polarity)	81% (34)	1.07	86% (42)	**2.31**
size	0	0.00	**18% (9)**	0.31
modality	2% (1)	0.02	**18% (9)**	0.22
moderator variable and qualifier**	24% (10)	0.29	**37% (18)**	0.71
mediator variable	7% (3)	0.10	**20% (10)**	0.35
direct or indirect relation	0	0.00	2% (1)	0.04
rebuttal	0	0.00	0	0.00
context	**43% (18)**	0.55	14% (7)	0.18
evidence	10% (4)	0.10	**22% (11)**	0.24
explanation	0	0.00	**14% (7)**	0.18
Comparison relation	7% (3 of 42)	0.19	**67% (33 of 49)**	2.06

Table 2

Percentage of Introduction sections containing each Research-relation element and Comparison relation, and average number per Introduction (N=50 Introduction sections)

	Research objective (N=46 Intro)		**Hypothesis (N=21 Intro)**		**Research question (N=15 Intro)**		**Research result (N=11 Intro)**	
Research-Relation Element	**% of Intro**	**Avg no. per Intro**	**% of Intro**	**Avg no.**	**% of Intro**	**Avg no.**	**% of Intro**	**Avg no.**
Research-relation (including concept1* and concept2)	98% (45 of 46)	3.96	89% (17 of 21)	2.52	87% (13 of 15)	3.13	82% (9 of 11)	4.45
• **cause-effect**	80% (37)	2.74	76% (16)	2.14	73% (11)	2.40	73% (8)	3.09
• **association**	33% (15)	1.13	*10% (2)*	0.38	33% (5)	0.73	36% (4)	1.36
• **prediction**	2% (1)	0.09	0	0.00	0	0.00	0	0.00
• **correlation**	0	0.00	0	0.00	0	0.00	0	0.00
instance (including polarity)	85% (39)	1.63	76% (16)	1.48	93% (14)	1.67	82% (9)	2.00
size	2% (1)	0.02	10% (2)	0.14	7% (1)	0.07	**27% (3)**	0.27
modality	2% (1)	0.02	**62% (13)**	0.81	13% (2)	0.27	**55% (6)**	0.64
moderator variable and qualifier	30% (14)	0.57	33% (7)	0.43	*7% (1)*	0.13	**45% (5)**	0.55
mediator variable	10% (5)	0.17	14% (3)	0.24	13% (2)	0.13	0	0.00
direct or indirect relation	0	0.00	5% (1)	0.05	7% (1)	0.07	0	0.00
rebuttal	0	0.00	5% (1)	0.05	0	0.00	0	0.00
context	**50% (23)**	0.72	10% (2)	0.01	20% (3)	0.20	36% (4)	0.27
evidence	**15% (7)**	0.22	0	0.00	7% (1)	0.07	0	0.00
explanation	9% (4)	0.09	**38% (8)**	0.48	0	0.00	18% (2)	0.09
Comparison relation	22% (10 of 46)	0.52	19% (4 of 21)	0.33	20% (3 of 15)	0.60	**64% (7 of 11)**	2.90

Research Objective

Context information has four subclasses: *location, environment, temporal* and *target population. Location* and *target population* often occur in the research objective statement in the Abstract and Introduction sections, to constrain the scope of the research study:

[Abstract] This study of $_{target\_population}$ *respondents from higher education institutions and research institutes examines the relationship ...*

[Introduction] This paper seeks to contribute … by examining the scientific system of an advancing country in Africa, namely location*South Africa.*

Environment and *temporal* can be seen as more detailed information provided in the Introduction section to limit the research scope. *Explanation* occurs more often in the Introduction section in the research objective, research hypothesis and research result statements. It plays different roles in these statements. In the research objective, it may be used as a justification for a Research-relation, for example:

We also examined whether the associations we observed can be explanation*explained by mothers' individual or family context or ...*

Hypothesis

Other than in a research result statement, *modality* and *polarity* are often indicated in research hypotheses in the Introduction section, for example:

It is modality*possible that an intention formed at a greater temporal distance reflects a stronger preference for a product and* modality*may therefore result ...*

Specifically, we hypothesize that weak tie use will polarity.positive*increase as a result of the shift from ...*

An underlying *explanation* is also often specified in research hypotheses:

Because explanation*neighborhoods are often ethnically homogeneous and because adolescents often attend schools nearby their homes, …*

A *Rebuttal* element derived from the literature may occur in a hypothesis statement in the Introduction section:

However, as Mau and Burkhardt (2009) argue, rebuttal*individuals' attitudes depend on distinct national socioeconomic and institutional contexts.*

Research Question

A research question generally has various types of information. It may have *polarity, modality, size, moderator* and *mediator* that specify a Research-relation in detail:

Do hotel websites with embedded social media channels have polarity.positive*higher levels of travelers' satisfaction?*

Third, do the mechanisms underlying the total effect of marriage vary across moderator*gender-race subgroups?*

… are the effects direct/indirect_relation*direct or* modality*largely mediated through* mediator*intervening experiences and exposures?*

Research Result

Modality, *polarity* and *size* are commonly used to specify research results, especially in the Abstracts. For example:

> *… and that growth in between-class income differences had a* $_{size}$*large* $_{polarity.positive}$*inflationary effect on trends in personal income inequality.*

A Rebuttal element may occur in a research result statement, especially in the context of a negative result. This example is from an Abstract:

> $_{rebuttal}$*The results do not support the role of coordination demands; the extent of...*

This example can also be coded as *modality.negation*.

Evidence has subclasses *research method, data source, information source* and *significance level*. *Research method* and *data source* are mainly found in research method statements. *Information source* and *significance level* are often indicated in result statements in the Abstract.

Moreover, a research result statement may suggest an underlying *explanation*:

> *As income inequality rises, middle-class identities become weaker…* $_{explanation}$ *because the adverse effects of inequality are felt more acutely across the class structure.*

Linking Comparison and Research-relation Frames

We found six link patterns between Comparison and Research-relation frames. About half the Abstracts (66%, 33 of 50 Abstracts) and Introduction sections (42%) have the Research-relation frame (the focal frame) linked to a Comparison frame. A comparison result/conclusion with a qualitative value was more common in the Abstract than in the Introduction section. This indicates that research results often involve a Comparison of a concept's attributes/aspects. The link patterns indicate different ways in which comparisons are used to support the argument claim that the Research-relation is valid. The six link patterns are illustrated in Figure 4 to 9.

Link Pattern 1: Comparison of Concept1 (Cause Concept) Subclasses/Attributes/Aspects

Comparison of two subclasses, attributes or aspects of *concept1* (the *cause* concept) based on their scores on some criterion attribute related to *concept2* (the *effect* concept) is commonly used to establish a Research-relation between *concept1* and *concept2* in a research result statement. The Comparison result can be an attribute value of *concept2,* or the *polarity* and/or *size* of the Research-relation. The following example illustrates this:

> *The results also indicate that* $_{Comp.subclass1}$*personal identification has a* $_{Comp.difference1}$*larger influence on* $_{Comp.measure}$*service brand loyalty than* $_{Comp.subclass2}$*social identification does.*

In a research question statement, there may be a comparison of subclasses, attributes or aspects of *concept1* based on the criterion measure, without, of course, giving the comparison result:

> *Are Comp.common_concept particular message formats, or ways of transmitting the information, more Comp.measure prone to error than others?*

In a research question, there is often an implied Comparison between the *cause* attribute category that is mentioned in the research question and the alternative category that is not mentioned but implied:

> *Do Comp.common_concept hotel websites Comp.attribute1 with embedded social media channels have Comp.difference1 higher levels of Comp.measure travelers' satisfaction, and do they improve travelers' purchase intentions?*

An implied Comparison may occur in the research result statement:

> *… the authors found that Comp.attribute1 higher levels of Comp.common_concept maternal education were associated with Comp.difference1 more advantageous Comp.measure health investment behaviors at each phase of ...*

There is an implied comparison between *higher* versus *lower levels* of maternal education. The comparison result indicates that *higher levels* is associated with the comparison result *more advantageous* (i.e. *difference1*), implying that *lower levels* is associated with the comparison result *less advantageous* (the *difference2* that is not explicitly mentioned). What is interesting is that the next sentence indicates a higher level comparison involving the moderator variable *developmental stage*.

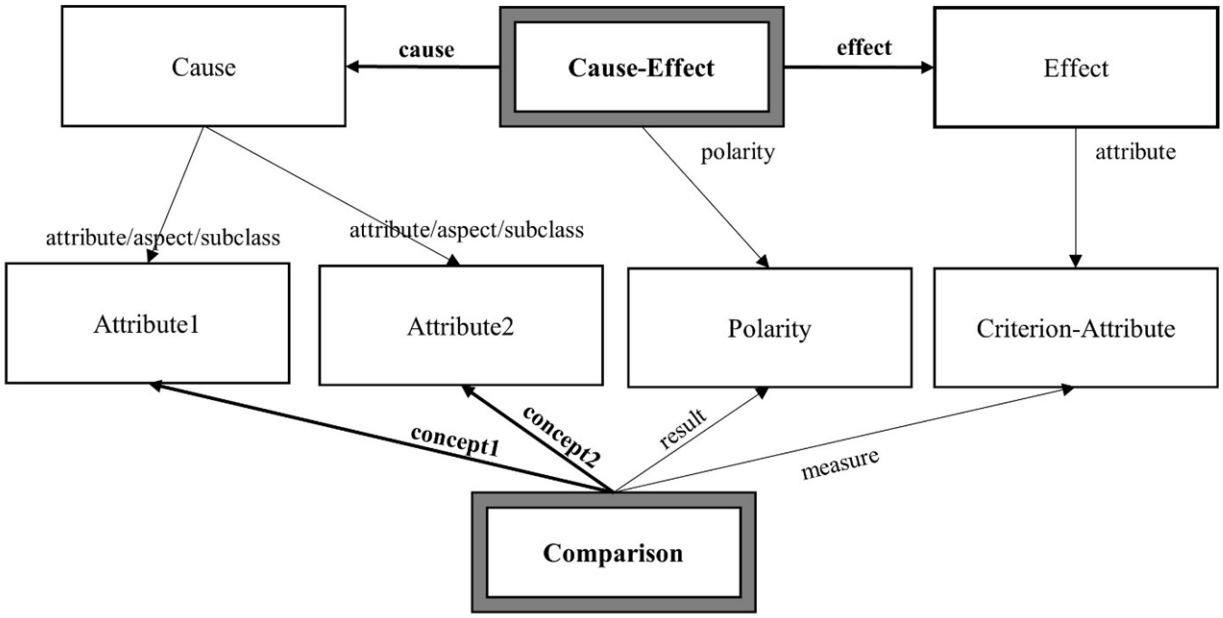
In research result statements, Comparison results are often used to infer a Research-relation, as well as different elements of the Research-relation. The qualitative value of a comparison result can be used to identify the *size* (magnitude) and *polarity* of the Research-relation:

> *I find that Research-relation.concept2 income segregation RR.polarity.positive increased only RR.concept1 **among families with children.** RR.concept1 **Among childless households**—two-thirds of the population—RR.concept2 income segregation RR.modality.negative changed little and is RR.size **half as large as among households with children**.*

Here, the Comparison is between attribute values of the *cause* concept (i.e. between *families with children* and *childless families*). The Comparison result is used to infer the *polarity* of the Research-relation between *families with children* and *income segregation*, as well as the relative *sizes* of the *income segregation*.

Figure 4

Link Pattern 1: Comparison of Cause Subclasses/Attributes/Aspects



Link Pattern 2: Comparison of Concept2 (Effect Concept) Subclasses, Attributes or Aspects

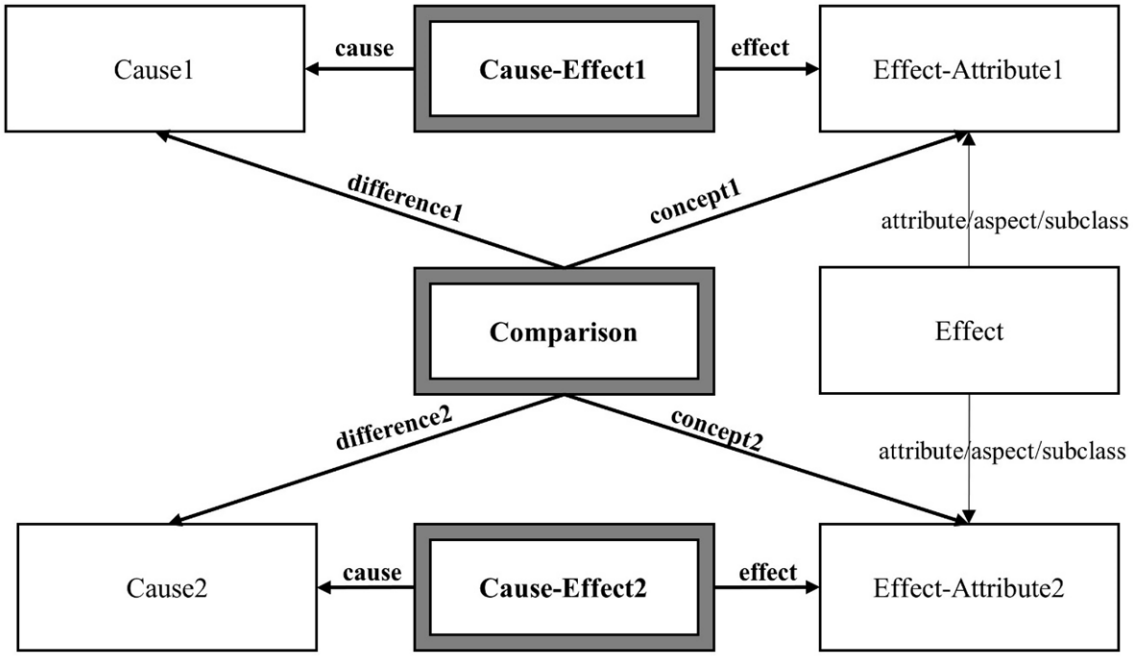
In research objective and result statements, there is occasionally a comparison of the *effect* concept's attributes/aspects/subclasses. The comparison result is that the two *effect* attributes have different *cause* concepts:

> *Comp.difference1*economic affluence and Jewish identity **predict** *Comp.attribute1*whiter *Comp.common_concept*self-identification, *Comp.instance*whereas *Comp.difference2*belonging to a religion more commonly associated with racial minorities *is* **associated with** *Comp.attribute2*a minority *Comp.common_concept*identification.

As the *cause* concepts in the Research-relation are quite different, we can consider this a comparison of two Research-relations.

Figure 5

Link Pattern 2: Comparison of Effect Subclasses/Attributes/Aspects



Link Pattern 3: Comparing Two Subclasses of Moderator and Mediator Variables

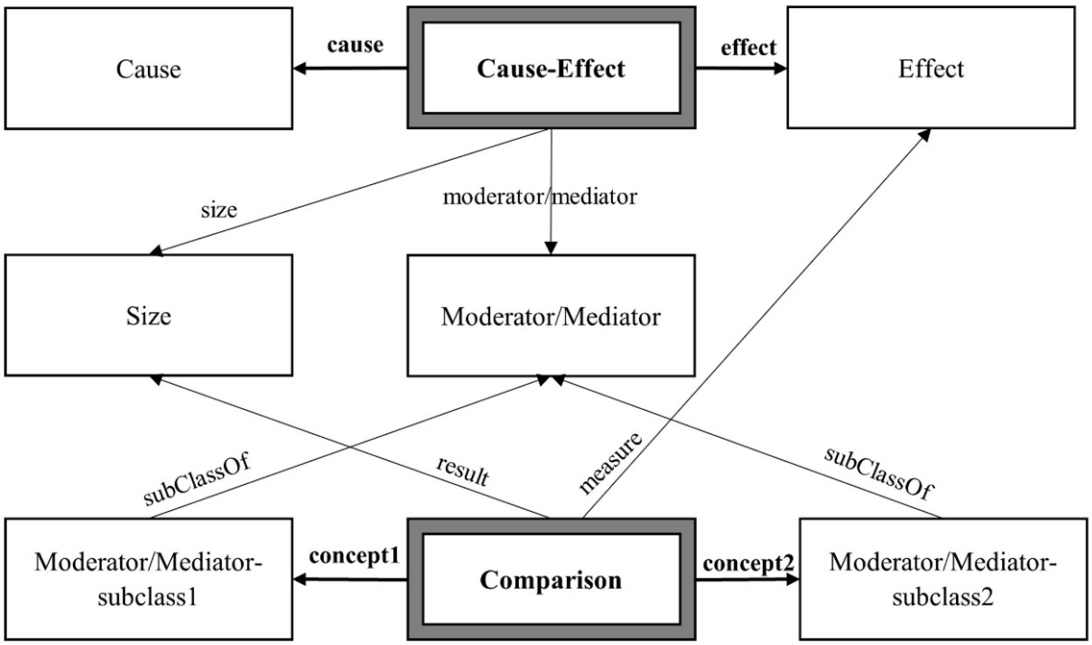
Research results are sometimes generated by comparing two subclasses of a moderator or, more rarely, a mediator variable. The *measure* and *comparison result* are often *concept2* (*effect* concept) of the Research-relation:

> *The results indicate that CSR-brand fit strengthens both RR.mediator*personal and social brand identification … The results also indicate that RR.mediator.subclass1*personal identification has a*

*Comp.difference1*larger *influence on* *Comp.measure*service brand loyalty *than* *RR.mediator.subclass2*social identification *does.*

Figure 6

Link Pattern 3: Comparing Subclasses of Moderator/Mediator Variables



Link Pattern 4: Comparing Different Underlying Explanations

In a research result statement, a comparison may link two Research-relation frames with related *cause* concepts (which may be subclasses of a broader concept), but is focused on highlighting different underlying explanations (the comparison result):

*Research-relation.common_concept*Mainland Chinese students *studying in Hong Kong* *RR.aspect1*actively use SNSs *for* *RR.explanation*seeking practical information about offline matters, and they obtain substantial enacted support from other Mainland students of the same university through SNS use. *RR.instance*As a result, *they* *RR.aspect2*accumulate both bridging and bonding social capital. *Research-relation,common_concept*Local Hong Kong students, *however,* *RR.explanation*use SNSs mainly for social information seeking *and are only able to* *RR.instance* accrue *RR.aspect2*limited bridging social capital *through* *RR.aspect1*SNS use.

Here, the comparison is between *Mainland Chinese students* and *Local Hong Kong students*, highlighting differences in the explanations (*difference1* and *difference2*), as well as differences in the effect of the causal relation.

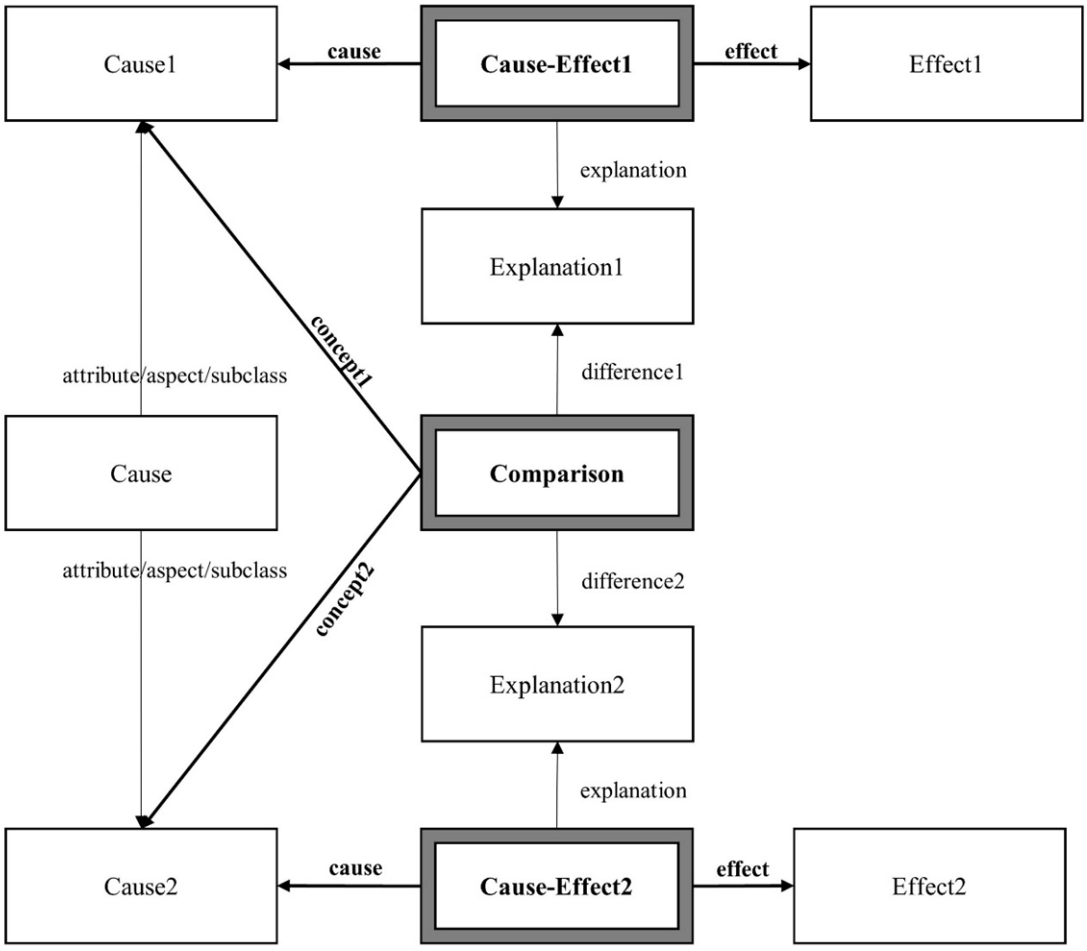
A comparison between explanations/theoretical mechanisms may also occur in research questions:

> *First, does the* Research-relation.concept2*wage* RR.instance*effect of* RR.concept1*marriage* RR.attribute2*take place instantaneously or cumulatively? Second, does* RR.attribute2*the life course pattern of the wage effect of marriage* ***vary by*** RR.moderator*race? Third, do* RR.explanation*the mechanisms underlying the total effect of marriage* ***vary across*** RR.moderator*gender-race subgroups?*

The keywords *vary by* and *vary across* indicate a comparison. *vary by* moderator*race* indicates comparison among the moderator (race) categories. "… the mechanisms underlying … *vary across* moderatorgender-race subgroups" indicates that the criterion attribute measured is "the mechanisms underlying".

Figure 7

Link Pattern 4: Comparing Different Underlying Explanations



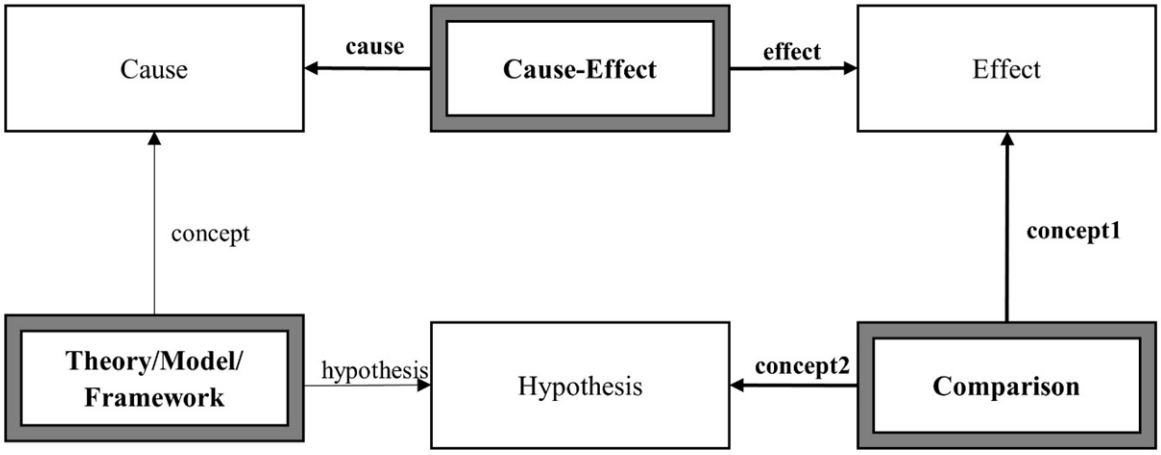
Link Pattern 5: Comparing Results to Expectation, Hypothesis or Theory

The research result statement may compare the study result regarding a Research-relation with commonsense expectation, or the result predicted by a hypothesis statement or derived from a theory:

Germany's contribution-based and highly work-oriented welfare state, and its historically...
Comp.instance ***In contrast to*** *Comp.concept1* ***this expectation,*** *Comp.concept2* ***our results*** *point to* *Comp.difference2* *a negative relationship between levels of the foreign-born population and German natives' attitude toward welfare support, which is both highly significant and distinctive.*

Figure 8

Link Pattern 5: Comparing Result to Expectation, Hypothesis or Theory



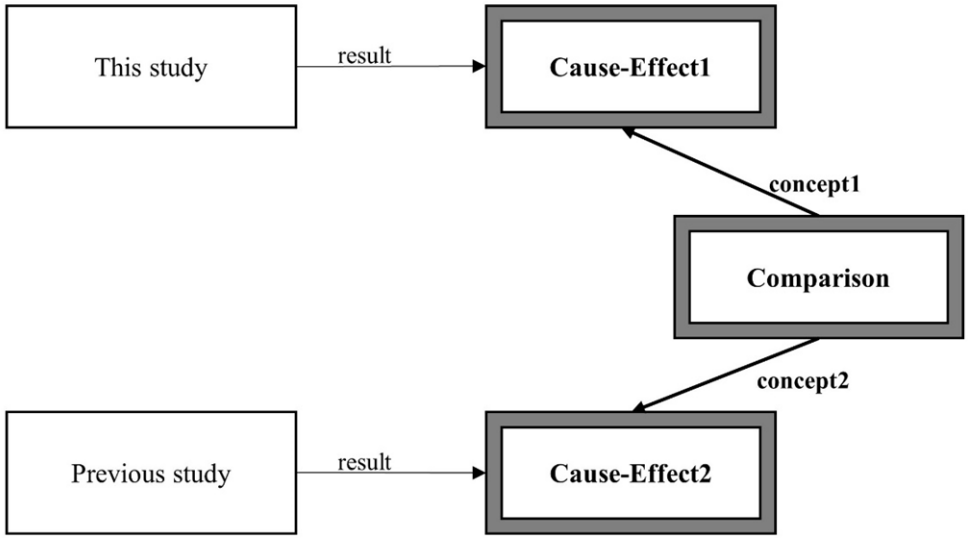
Link Pattern 6: Comparing Results to Those of Previous Study

The research result statement may contrast the study results regarding a Research-relation with the results of one or more previous studies:

Comp.instance *In contrast to a recent cross-country* *Comp.concept1* *study by Brady and Finnigan (2014),* *Comp.concept2* *we conclude that the relationship between migration and concerns about welfare is not restricted to the United States but can also be observed in the very unlikely case of Germany .*

Figure 9

Link Pattern 6: Comparing Result to Those of Previous Study



Conclusion

We have introduced a method of information structure analysis of academic text using semantic frames, including analyzing how the frames are linked together to support research arguments. We developed two semantic frames, the Research-relation frame and the Comparison frame, which specify the types of information associated with *cause-effect* relations and *comparison* relations, which are expected to be expressed in the text.

We used the semantic frames as coding schemes to analyze 50 sociology abstracts and introduction sections—to identify the characteristic information profiles of research objective, hypothesis, research question and research result statements. Not surprisingly, research result statements, tend to contain more details (i.e. elements of the Research-relation frame). However, research objective statements are more likely to carry *context* information. *Moderator/qualifier* variables are more often found in research result statements. Comparisons are found in about two-thirds of research result statements, but only in 20% of research objective, hypothesis and research question statements.

About one-third of the research objective, research question and research result statements contain an *association* relation (which is weaker than a *cause-effect* relation). The *association* relation is seldom found in hypothesis statements, which usually hypothesize the stronger *cause-effect* relation. Hypothesis statements are more likely to include underlying *explanations* because hypotheses are often

derived from theory. *Modality* is more likely to be found in the hypothesis and research result statements, usually indicated by modal adverbs (e.g., *may*, *significant*).

We also identified six link patterns between Comparison and Research-relation frames, showing how comparisons are used to support the argument claim that the Research-relation (including cause-effect) is valid: comparing subclasses/attributes/aspects of the *cause* concept or the *effect* concept; comparing subclasses of *moderator* or *mediator* variables; comparing possible underlying explanations for a Research-relation; and comparing results to expectation, hypothesis, theory or a previous study's result. Through examples, we examined the details of how the comparison relation supports the cause-effect relation.

Information structure analysis can identify more detailed conceptual structure underlying argument structure and argumentation schemes. Indeed, our link patterns between Comparison and Research-relation frames bear resemblance to Green's (2015) argumentation schemes, especially the schemes that involve comparison of observation (research result) with expected results (hypothesis). The major difference is that Green's argumentation schemes are focused on logical reasoning based on sets of research subjects, whereas quantitative research in sociology focuses on statistical results and probabilities. Also, our analysis is of the Abstract and Introduction sections of papers. It is possible that Green's argumentation schemes are used more frequently in the Results and Discussion sections of sociology research papers.

Further research is warranted to explore the relation between information structure and other types of arguments. In follow-up work, we identified several link patterns between Cause-effect and Association frames (subclasses of the Research-relation frame) in the Research gap=>Research objective and Research hypothesis=>Research objective argument steps, showing how Research-relations are specialized or generalized from a Research gap/hypothesis to the Research objective (Cheng, 2020). We have also analyzed link patterns between the Theory/model/framework frame and Research-relation frame, which occurs in about 67% of the Introduction sections in our corpus.

This study is limited to papers reporting *Investigative research* that investigates cause-effect relationships. In follow-up work, we have extended the analysis to other types of sociology research: *Development and evaluation research*, and *Descriptive research*. The study is also limited to the abstract and introduction sections. It should be extended to other sections of research papers. Information structures in research papers can be quite different in different disciplines, and so, comparative studies should be carried out in different disciplines. Of course, a major challenge is to then develop an automated method to parse the information structure of text to extract information to populate social science research knowledge graphs.

Acknowledgement

This study was funded partly by the Singapore Ministry of Education research grant MOE2015-1-TR05.

References

Ahlstrom, D. (2017). How to publish in academic journals: Writing a strong and organized introduction section. *Journal of Eastern European and Central Asian Research*, 4(2), 1-9. **doi: 10.15549/jeecar.v4i2.180**

Brack, A., Hoppe, A., Stocker, M., Auer, S., & Ewerth, R. (2020). Requirements analysis for an open research knowledge graph. In *24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020* (Lecture Notes in Computer Science, vol 12246, pp. 3-18). Springer Nature. **https://doi.org/10.1007/978-3-030-54956-5_1**

Burbules, N. C. (2015). The changing functions of citation: From knowledge networking to academic cash-value. *Paedagogica Historica, 51*(6), 716-726. **https://doi.org/10.1080/00309230.2015.1051553**

Cheng, W. -N. (2020). Argument and information structures in sociology research papers: Analysis of the Abstract and Introduction sections. Doctoral thesis, Nanyang Technological University, Singapore. **https://hdl.handle.net/10356/138530**

Ehrlinger, L., & Wöß, W. (2016). Towards a definition of knowledge graphs. In *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16).* **https://dblp.org/db/conf/i-semantics/semantics2016p.html**

Fillmore, C. J. (1968). The case for case. In E. Bach, & R. T. Harms (Eds.), Universals in linguistic theory (pp. 1-88). New York: Holt, Rinehart, and Winston.

Fillmore, C.J., Johnson, C. R., & Petruck, M. R. L. (2003). Background to FrameNet. *International Journal of Lexicography*, *16*(3), 235–250. **https://doi.org/10.1093/ijl/16.3.235**

Flowerdew, J. (1999). Problems in writing for scholarly publication in English: The case of Hong Kong. *Journal of Second Language Writing, 8,* 243-264. **doi:10.1016/S1060-3743(99)80116-7**

Gábor, K., Buscaldi, D., Schumann, A. K., QasemiZadeh, B., Zargayouna, H., & Charnois, T. (2018). SemEval-2018 Task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of the 12th International Workshop on Semantic Evaluation, SemEval-2018* (pp. 679-688). Association for Computational Linguistics. **https://www.aclweb.org/anthology/S18-1111/**

Gee, J. P. (2014). An introduction to discourse analysis: Theory and method (4th ed.). Routledge.

Green, N. L. (2015). Identifying argumentation schemes in genetics research articles. In C. Cardie (Ed.), *Proceedings of the 2nd Workshop on Argumentation Mining*, Denver, CO (p. 12–21). The Association for Computational Linguistics.

Gutierrez, C., & Sequeda, J. F. (2019). *A brief history of knowledge graph's main ideas: A tutorial*. **http://knowledgegraph.today/paper.html**

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.

Jaradeh, M. Y., Oelen, A., Farfar, K.E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M., and Auer, S. (2019). Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture (K-CAP '19).* ACM. **https://doi.org/10.1145/3360901.3364435**

Johnstone, B. (2017). *Discourse analysis* (3rd ed.). Wiley-Blackwell.

Kwan, B. S. C., Chan, H., & Lam, C. (2012). Evaluating prior scholarship in literature reviews of research articles: A comparative study of practices in two research paradigms. *English for Specific Purposes, 31* (3), 188-201.

Lovejoy, K. B. (1991). Cohesion and information strategies in academic writing: Analysis of passages in three disciplines. *Linguistics and Education*, *3*(4), 315–343.

Lim, J. M. -H. (2011). 'Paving the way for research findings': Writers' rhetorical choices in education and applied linguistics. *Discourse Studies*, *13*(6), 725–749.

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*, 797–813. **https://doi.org/10.1007/s11192-018-2770-2**

Lincoln, Y. S., Lynham, S. A., & Guba, E. G. (2011). Paradigmatic controversies, contradictions, and emerging confluences, revisited. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (4th ed., pp. 97-128). Sage Publications.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text: Interdisciplinary Journal for the Study of Discourse, 8*, 243–281. **https://doi.org/10.1515/text.1.1988.8.3.243**

Mann, W. C., & Taboada, M. (2021). *Intro to RST (Rhetorical Structure Theory)* [web page]. http://www.sfu.ca/rst/index.html

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, *27*(1), 57–86. **https://doi.org/10.1177/0741088309351547**

Miles, B. (2010). Discourse analysis. In N. J. Salkind (Ed.), *Encyclopedia of Research Design* (pp. 368-370). SAGE Reference.

Ou, S., Khoo, C. S. G., & Goh, D. (2007). Automatic multi-document summarization of research abstracts: Design and user evaluation. *Journal of the American Society for Information Science & Technology*, *58*(10), 1419–1435. **https://doi.org/10.1002/asi.20618**

Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics & Pragmatics*, *5*, article 6. **http://dx.doi.org/10.3765/sp.5.6**

Schwandt, T. A. (1998). Constructivist, interpretivist approaches to human inquiry. In N. K. Denzin & Y. S. Lincoln (Eds.), *The landscape of qualitative research: Theories and issues* (pp. 221-259). Sage Publications.

Schiffrin, D., Tannen, D., & Hamilton, H.E. (2001). Introduction. In D. Schiffrin, D. Tannen, & H.E. Hamilton (Eds.), *The Handbook of Discourse Analysis* (pp. 1). John Wiley & Sons.

Slaughter, L., Berntsen, C. F., Brandt, L., & Mavergames, C. (2015). Enabling living systematic reviews and clinical guidelines through semantic technologies. *D-Lib Magazine, 21*(1/2). **https://doi.org/10.1045/january2015-slaughter**

Stremersch, S., Camacho, N., Vanneste, S., & Verniers, I. (2015). Unraveling scientific impact: Citation types in marketing journals. *International Journal of Research in Marketing*, *32*(1), 64–77. **https://doi.org/10.1016/j.ijresmar.2014.09.004**

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

Zhou, D., Zhong, D., & He, Y. (2014). Biomedical relation extraction: From binary to complex. *Computational and Mathematical Methods in Medicine, 2014*, article ID 298473. **http://dx.doi.org/10.1155/2014/298473**