The Many Uses of Explain:

Quantitative Corpus Method and Philosophy of Science

Yuki SUGAWARA* and Kazuho KAMBARA**

Abstract

Recent studies in the philosophy of science have employed data science methods, which have the potential to overcome the limitations of conventional case study approaches. These limitations include a lack of interest in the typicality of language uses and arbitrary case selection. The digital philosophy of science has implemented text mining and random sampling techniques. This paper aims to address methodological issues in the digital philosophy of science and argue for refining quantitative concept analysis. Specifically, we focus on the various uses of the word *explain* to demonstrate the effectiveness of this approach. By integrating methodologies from the philosophy of science and contemporary linguistics, we propose an updated approach to the digital philosophy of science.

Key words: Naturalism, Explanation, Digital Philosophy of Science, (Quantitative) Concept Analysis, QCM (Quantitative Corpus Method), Frame Semantics, Ontology and Epistemology

1. Introduction

This paper aims to demonstrate the effectiveness of corpus-based quantitative concept analysis through a detailed analysis of *explain*. This approach utilises various data science methods for concept analysis in the philosophy of science. Specifically, we utilise the Quantitative Corpus Methods (QCM) [18], a set of corpus linguistics

* Osaka University
E-mail: ysugawara.hmt@osaka-u.ac.jp
** Ritsumeikan University
E-mail: kazy0324@pep-rg.jp
Portions of this study were reported at the 48th conference of Japanese Association for English Corpus Studies (JAECS).

methods, to argue for the need to incorporate linguistic methodologies in quantitative concept analysis.

Naturalistic perspectives of philosophy have been prevalent in the philosophy of science since the early 20th century [33]. Some naturalists argue for a continuous relationship between philosophy and science [5, 6, 7]. However, despite recent developments in linguistics and data science, only a few philosophers have applied these state-of-the-art frameworks in their linguistic analyses. This paper proposes integrating modern linguistic and data science methods in concept analysis, providing an alternative perspective on the naturalisation of the philosophy of science.

The paper is structured as follows. §2 discusses the limitations and characteristics of recent quantitative concept analysis and introduces the Quantitative Corpus Methods (QCM). §3 outlines our research methods and annotation strategies based on cognitive semantics. §4 presents our data analysis of the verb *explain*. Finally, §5 discusses the implications of our findings. The paper concludes in §6 with a summary of our contributions and possible future developments.

2. Background: A path from concept to text

This section delves into the necessity of text analysis in the philosophy of science. §2.1 provides an overview of the quantitative shift in concept analyses within the philosophy of science and scrutinises their limitations. Subsequently, §2.2.1 elucidates significant features of QCM. Finally, §2.2.2 deliberates on the concordance of QCM with the digital philosophy of science.

2.1. Quantitative Analysis of Concepts in Philosophy of Science

This section examines the emergence of quantitative analyses in the philosophy of science. §2.1.1 provides an overview of the recent developments of quantitative methods in analysing concepts, while §2.1.2 critically discusses some of the limitations posed by previous studies.

2.1.1. Overview

This section provides an introduction to the recent developments in the field of **digital philosophy of science**. Over the last few years, philosophers of science have increasingly employed quantitative methods in their studies, commonly referred to as digital philosophy of science [32]. This approach builds upon the computational philosophy of science that emerged in the 1980s [37], utilising computational methods to investigate topics in the philosophy of science [2, 38].

Most philosophers of science traditionally rely on the **case study methods** [23]. According to Machery [23], philosophers of science have traditionally conducted research using the case study method, based on significant literature (such as textbooks

and papers) in specific scientific fields of interest, to gain a detailed understanding of the particular scientific practice, considering scientific knowledge and the history of science. Though the case study method is an effective approach based on a detailed understanding of the case, Machery point out that the method has two limitations: (i) the inability to capture differences in scientific research tied to specific cases, and (ii) exposure to bias during case selection. Machery argue that these problems with the case study method could be resolved through the **experimental philosophy of science**. Specifically, this would involve recording, and analysing scientists' language use by extracting data about specific language use from large corpora and scholarly literature. By collecting and analysing data on how scientists use language in a way that is not limited to specific cases, it is possible to avoid the two problems posed by the case study method.

In this respect, in recent years, there has been a gradual increase in the number of philosophy of science studies conducted using language data. The quantitative approaches for textual analysis have been increasingly popular among humanities, with the qualitative approach being termed "closed reading" and the quantitative approach "distant reading" [29]. Since the 2000s, some studies have attempted to employ distant reading in the philosophy of science [1, 24, 40, 41], providing new insights and specific examinations from a quantitative perspective.

However, the approach is not immune to criticism. First, philosophers of science tend to focus on "concepts" while ignoring the actual uses of "words" [3]. Since concepts only arise by generalising the actual language uses, it is insufficient just to "describe" concepts. Second, the case study method often needs to consider the typicality of the selected cases. To address these issues, Overton [31] employed text mining and random sampling in his study. By analysing the verb "explain" and semantically related words (e.g., *because*, *show*), Overton reframed the question of "what is the concept of *explain*" to "how is the concept of *explain* used in scientific practices". Additionally, he randomly sampled sentences including the words related to "explain" to overcome the "armchair" style of philosophical enquiries, which can be viewed as experimental philosophy.

In the digital philosophy of science, philosophers conduct experiments on texts in scientific practices. Texts are not arbitrarily chosen but are randomly sampled to ensure that they are treated as "natural" objects, similar to other natural objects such as trees and animals. By treating texts as natural objects, philosophers can employ scientific methods to generate and verify hypotheses as part of their research.

2.1.2. Limitations

This section highlights the limitations of the contemporary digital philosophy of science. As mentioned earlier, Overton [31] argued that quantitative studies have introduced (i) text mining and (ii) random sampling to resolve the problems associated

with case study methods. However, it is still debatable whether Overton's methods are appropriate for philosophical studies. Therefore, this section critically evaluates the methodological challenges in the contemporary digital philosophy of science, particularly focusing on Overton's research [31].

First, we assess the issue of text mining. Overton computed the frequencies of particular words, but a more sophisticated approach can be taken. Philosophical studies deal with concepts that have multiple subtypes with a hierarchical structure, often being neatly classified. Therefore, when digital philosophers quantitatively analyse concepts instead of words, they need to consider the frequencies of word meanings rather than the frequencies of particular words. In this regard, cognitive semantics can offer theoretical frameworks and practical procedures for text analysis. By performing linguistic annotations on texts and analysing the frequencies of meanings of words, digital philosophers can combine qualitative understanding in linguistics with quantitative understanding.

The second limitation concerns the method of random sampling. Overton selected specific words from a particular journal within a certain period, and the validity of his sampling method is debatable. In corpus linguistics, **corpus-driven** and **corpus-based** approaches are often distinguished [25, p.5–6]. Advocates of the corpus-driven approach attempt to analyse language (hopefully) without theoretical premises, while those of the corpus-based approach with theoretical premises and motivations. Often in a corpus-driven approach, analysts create their original dataset to investigate the phenomena of their choice. However, when creating such datasets, the typicality of their data is open to discussion. For instance, if an analyst decides to create a corpus of scientific journals like Overton, it is clear that the contents of the corpus are quite different from the "ordinal" uses of language. To reveal the typical uses of *explain* to contrast them with specific ones (i.e., "explain" in scientific discourse), analysts must employ a more balanced corpus, such as British National Corpus (BNC, for short). Though Overton's method radically changed the course of the philosophy of science, the typicality of his results is still open to discussion.

2.2. An alternative approach to text analysis

As discussed in the previous section, the conventional methods cannot capture the "conceptual" nature of the linguistic text. As an alternative, §2.2.1 introduces the quantitative corpus methods (QCM), and §2.2.2 argues that QCM can be seen as a finer-grained method of concept analysis.

2.2.1. Quantitative corpus method (QCM)

This section introduces some of the important characteristics of the **quantitative corpus methods** (QCM for short) [18] through a brief examination of the polysemous nature of *run*.

QCM is a set of analytical methods developed mainly in linguistics, which focuses on the quantitative evaluation of (manual and automated) analyses of linguistic text. Like philosophers of science, linguists have been eager to incorporate quantitative methods into their toolbox. The shift to quantitative methods is especially prominent in the recent development of **cognitive linguistics** [21]. Quantitative turn in cognitive linguistics emphasises the importance of experimental and corpus analysis in their research programme. QCM is a method derived from such a movement.

According to the survey conducted by Gries [18], monofactorial analysis (usually taught in an introductory course of statistics) is not suitable for most of the linguistic analysis since almost all of the linguistic phenomena is multifactorial[1]. Take one of the most well-known polysemous verbs, *run* [15, 16, 19]. It is impossible to distinguish the different senses (namely, FAST PEDESTRIAN MOVEMENT, ROMANTIC ESCAPE and BUSINESS MANAGEMENT, etc.) in (1) unless one considers the contextual features of the word. Even in this simplified demonstration, one needs to come up with multiple variables to distinguish these senses (e.g., transitivity of the verb, cooccurrence of adjuncts).

(1) a. Alice ran.
b. Alice ran with Bill.
c. Alice ran the restaurant.

As the name suggests, advocates of QCM commit to quantitative analysis of corpus data. One could use a corpus as a collection of linguistic data, which provides valuable data that an analyst cannot imagine otherwise [11]. In addition to a manual assessment of the given data set, one could employ quantitative analyses to reveal interesting aspects of language use. Gries demonstrated how different variables interact by combining manual annotations of texts and statistical analysis to show different senses of *run* [15]. Such analysis is only possible when an analyst conducts a consistent assessment of the data.

2.2.2. QCM as a friend of philosophy of science

This section points out that similar motivations for the quantitative method are observed in linguistics and science philosophy. As far as concept analysis in philosophy is based on linguistic analysis, QCM can be taken as a step toward the naturalisation of the philosophy of science [33].

The motivation of QCM arose due to the excessive exploitation of introspection in linguistic analysis. As discussed in §2.1, this is true in linguistics and the philosophy of science. This similarity is caused by the fact that both approaches describe language use. While philosophers of science aim to reveal the characteristics of an

[1] For a brief, yet broad introduction to statistical linguistic analysis, see [17]

abstract concept (e.g., CAUSATION, MECHANISM) which drives the various aspects of scientific discourse, linguists describe how variables introduced in the literature contributes to each use of words. For instance, a brief analysis of *run* discussed in §2.2.1 dealt with the behaviour of response variables (i.e., senses of *run*, namely: FAST PEDESTRIAN MOVEMENT, ROMANTIC ESCAPE and BUSINESS MANAGEMENT) in relation to various predictor variables (e.g., transitivity of the verb, cooccurrence of adjuncts). The targets of studies can vary depending on the motivations. However, as discussed in §2.1.2, the theory-neutral method of linguistic analysis falls short of a finer-grained analysis of the given concept.

In discussing the relation between scientific enquiries and philosophical investigations, naturalists (in the philosophical sense) emphasise the continuity between two disciplines [33, 34]. One of the most well-known naturalists, Dennett especially emphasises the importance of scientific methods in philosophical analyses [5, Ch.3]. For instance, to construct an adequate model of minds, Dennett dealt with results of various fields: evolutionary biology, cognitive psychology, computer science, etc. This methodology is tantamount to saying goodbye to good-old introspective enquiries.

As far as linguistic analysis is involved as a method of concept analysis, employing a linguistically-informed method is a natural consequence of philosophical naturalism. Though targets of analyses in linguistics are often quite different from those of philosophy of science, some of the analytical frameworks are still applicable to philosophically motivated linguistic analysis. For this reason, we aim to demonstrate how a linguistically-informed method can be applied to the conceptual analysis of *explain*. This way, QCM can be treated as a friend (rather than a foe) of the philosophy of science.

However, committing to the methodology of linguistics comes with a few caveats. The first is that the linguistically-informed quantitative concept analysis inherits the strength and limitations of the current method. The second is that the boundary between philosophical concept analysis and quantitative analysis blurs. Let us call the first issue as **the problem of blind inheritance**, and the second as **the problem of blurry boundary**.

Without a doubt, QCM offers appealing approaches to perform fine-grained linguistic analysis and has proven valuable in the field of lexical semantics (especially the studies of semantic relations like synonymy and antonymy[2]). Despite its appeal-

[2] In lexical semantics, semantic relations between words are broadly characterised into four categories [4, 30], namely: (i) synonymy (based on the similarity of meaning [e.g., *sofa–couch*]), (ii) antonymy (based on the oppositeness of meaning [e.g., *tall–short*]), (iii) hyponymy (based on the inclusion of meaning [e.g., *dog–animal*]), and (iv) meronymy (based on the part-whole relation of the meaning [e.g., *finger–hand*]). Though the quantitative method has proven useful, especially in the study of synonymy and antonymy, its effectiveness in the study of hyponymy and meronymy is yet to be explored.

ing outlook, linguistic semantics has centred its target of analysis around predicative expressions like verbs and adjectives. As discussed above, philosophers tend to analyse abstract concepts regardless of the word classes. For instance, the difficulty in analysing the noun *causation* would not decrease even when one employs QCM.

Even after resolving the methodological puzzle posed by the problem of blind inheritance, one has to struggle with the blurry boundary. When a philosopher and a linguist decide to analyse the same word, it becomes difficult to distinguish one enquiry from another. For example, the one conducted by Overton [31] can also be seen as a lexicographic analysis of *explain*. All QCM studies do not necessarily contribute to the development of philosophy and vice versa.

These criticisms are not new to the naturalisation of philosophy. Firstly, The problem of blind inheritance presupposes the Platonistic view of scientific methodology because it assumes the existence of a superior analytical framework while ignoring the benefits of the currently available methodology. As finite beings, we have no choice but to accept the tools available to answer the questions at hand. Secondly, the problem of blurry boundaries is not a problem when each research is motivated by different factors. In any empirical research, the target of the study is selected for a reason, typically derived from theoretical predictions or research questions. As far as different motivations are involved, the problem of blurry boundary is nothing but an illusion[3].

Despite the possible criticisms, QCM still holds the hope to hone the traditional concept analysis. Regarding the quantitative method in concept analysis, it is not realistic to ignore the benefits of QCM in concept analysis. As discussed in §2.1, Overton's approach to *explain* has a lot to improve from the standpoint of QCM. For this reason, we perform a QCM analysis of *explain* to compare with Overton's analysis.

3. Method

This section introduces the methodological strategies and statistical analyses employed in this study. This section is structured as follows: §3.1 introduces the strategy employed in this paper, §3.2 explains the statistical techniques to process the data, and §3.3 describes the procedure.

3.1. Annotation Strategies

This section introduces strategies employed in this study. §3.1.1 explains the strategy for semantic information, and §3.1.2 the strategy for other information.

[3] Moreover, one can argue that an analyst has the freedom to interpret the result of analyses following her research tradition. In this paper, we will not pursue this path.

3.1.1. Semantic Annotation

This section introduces basic assumptions in the semantic annotation. We employed **frame semantics** to record semantic features of examples involving the verb *explain*. In the following, we overview the essential characteristics of frame semantics to explicate the semantic annotation procedure taken in this study.

In frame semantics, a word meaning is described relative to a (typically) situational concept, called a **frame** [13, 14]. Generally, a frame is characterised as a set of semantic roles and their relations among them[4]. When a word's meaning is based on a specific frame, the word meaning is said to **evoke** a frame [13, p.236]. For example, the verb *buy* evokes `Commercial transaction`, which specifies a complex interaction among semantic roles such as ⟨BUYER⟩, ⟨SELLER⟩, ⟨GOOD⟩, and ⟨MONEY⟩. We refer to semantic roles specific to a frame as **frame elements**. By employing frame semantics, one can coherently analyse some verbs, as (2).

(2) a. Alice ⟨BUYER⟩ **bought** the car ⟨GOODS⟩ from Bill ⟨SELLER⟩.
b. Bill ⟨SELLER⟩ **sold** the car ⟨GOODS⟩ to Alice ⟨BUYER⟩.
c. Alice ⟨BUYER⟩ **paid** $10,000 ⟨MONEY⟩ for the car ⟨GOODS⟩.
d. Bill ⟨SELLER⟩ **charged** $10,000 ⟨MONEY⟩ for the car ⟨GOODS⟩.

To conduct a frame semantic annotation, we devised `Explaining` and its frame elements as defined in (3–4). Though philosophers of science tend to focus on the scientific discourse, many usages of the verb *explain* do not necessarily represent "scientific" contents. For this reason, we assumed two sub-classes of `Explaining` which subsumes (i) $\texttt{Explaining}_1$ representing an "ordinal" discourse, and (ii) $\texttt{Explaining}_2$ representing an "scientific" discourse. Descriptions in (3–4) correspond to a coherent, general scenario called "explanation".

(3) `Explaining`:
a. $\texttt{Explaining}_1$: ⟨EXPLAINER⟩ conveys ⟨TOPIC⟩ (in some occasion to ⟨AUDIENCE⟩) with or without ⟨MEDIUM⟩
b. $\texttt{Explaining}_2$: ⟨EXPLANAN⟩ captures some characteristics of ⟨EXPLANANDUM⟩, often being phenomena

(4) a. Frame elements of $\texttt{Explaining}_1$:
i. ⟨EXPLAINER⟩: An agent who attempts to convey ⟨TOPIC⟩ to ⟨AUDIENCE⟩ (abbreviated as Exp).

[4] It is useful to assume both dynamic and static relations for frames. For instance, `Human body` can be analysed in terms of its typical body parts, such as ⟨TRUNK⟩, ⟨ARM⟩, etc. However, for simplicity, we only deal with a dynamic relation like `Commercial transaction`.

ii. ⟨Topic⟩: A content conveyed to ⟨Audience⟩ by ⟨Explainer⟩ (abbreviated as Top).

iii. ⟨Audience⟩: A person (or, people) who receives a conveyance of ⟨Explainer⟩ (abbreviated as Aud).

iv. ⟨Medium⟩: A medium that is used by ⟨Explainer⟩ to effectively express ⟨Topic⟩ (abbreviated as Med).

b. Frame elements of `Explaining`$_2$`:`

i. ⟨Explanan⟩: A theoretical construct to capture some characteristics of ⟨Explanandum⟩ (abbreviated as En).

ii. ⟨Explanandum⟩: An entity or phenomenon that is investigated (abbreviated as Ed).

Defining a frame and its corresponding frame elements is insufficient to reveal the characteristics of *epxlain*. Frame semantics enables analysts to describe the relationship between a frame and corresponding expressions. Frame semantic annotation of a text reveals how a frame is realised linguistically, which provides "deep" semantics [12] that cannot be obtained from surface structures of texts. For instance, annotated sentences in (2) represent how `Commercial transaction` is realised by conjoining realised frame elements. We refer to the array of realised frame elements as **subarray** of a frame.

3.1.2. Other Annotation

This section introduces the grammatical features used in this study. Frame semantic annotation can reveal how the frame is realised, but it does not itself tell us what kind of (para-)linguistic factors are involved. For this reason, we employed (i) grammatical features and (ii) genre features. Grammatical features are recorded to investigate how each subarray is realised linguistically, and genre features to the kind of subarrays that are likely to be used in each text genre. The grammatical features we employed are defined in (5), which were obtained in a bottom-up fashion since it is difficult to define these features in a priori fashion. For genres, we employed pre-annotated information as explicated in (6).

(5) Grammatical features:

a. isPassive: 1 if the sentence in question is in passive voice, 0 otherwise.

b. Modality: 1 if the sentence in question is in imperative, 0 otherwise.

c. Quotation: 1 if the sentence in question involves quotation, 0 otherwise.

d. Explicit ⟨Explainer⟩: 1 if the sentence in question realizes ⟨Explainer⟩, 0 otherwise.

(6) Genre:

Applied science, Arts, Belief & thought, Commerce & finance, Imaginative, Leisure, Natural & pure science, Social science, Unknown, World affairs

3.2. Statistical analysis

This section introduces the employed statistical techniques: (i) Cluster analysis (§3.2.1), (ii) correspondence analysis (§3.2.2), and (iii) logistic regression analysis (§3.2.3).

3.2.1. Cluster analysis

In this study, we used cluster analysis to discover structures in linguistic data. As Divjak and Fieller describe in their introductory chapter on cluster analysis for linguistic data [9], cluster analysis is "an exploratory data analysis technique, encompassing a number of different algorithms and methods for sorting objects into groups" [9, p.405]. In cluster analysis, for example, as shown in Table 1 [9, p.419], numerical variables for each language's characteristics are used as the basis to calculate the similarity between each language as a distance [9, p.420]. This distance is then used to group them [9, p.423]. Similarly, in this study, frequency tables for each subarray of `Explaining` were created, and hierarchical cluster analysis was conducted based on these tables.

In hierarchical cluster analysis, we process the given data by (1) sequentially grouping the most similar data, and (2) forming collections with a significant hierarchical structure from the groups created in (1). The criterion for grouping data in (1) is referred to as **inter-individual dissimilarity**, and for grouping the clusters in (2) is called **inter-cluster dissimilarity**. It is necessary to define these two measures of dissimilarity to perform hierarchical cluster analysis.

Inter-individual dissimilarity is a metric used to group objects to form clusters. Examples of this measure include Euclidean distance, Manhattan distance, Chebyshev distance, Canberra distance, Binary distance, and Minkowski distance. Inter-cluster dissimilarity, on the other hand, is the criterion used to group clusters

Table 1 An example of three numerical variables measured on six languages [9, p.419]

Language	Number of letters in the alphabet	Number of speakers in billions	Official language in number of countries
English	26	0.360	7
German	30	0.120	3
Dutch	26	0.028	6
Russian	33	0.155	5
Polish	32	0.040	1
Serbian	30	0.0087	3

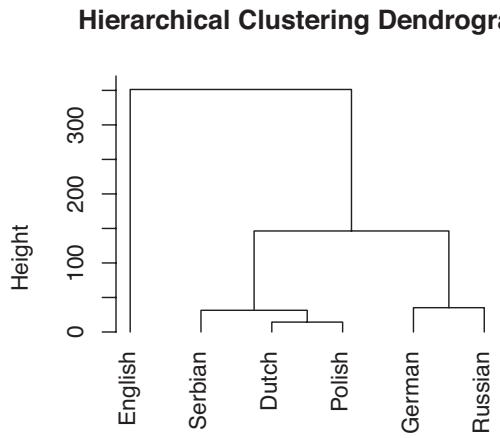


Figure 1 An example of cluster analysis

Table 2 An example of the distance matrix of Euclidean distances between languages [9, p.420]

	English	German	Dutch	Russian	Polish	Serbian
English	0	0.5662	1.054	7.283	8.491	5.668
German	5.662	0	5.001	3.606	2.830	0.111
Dutch	1.054	5.001	0	8.127	11.012	7.019
Russian	7.283	3.606	8.127	0	5.15	5.146
Polish	8.491	2.830	11.012	5.115	0	4.031
Serbian	5.668	0.111	7.019	5.146	4.031	0

together. This can be based on different methods like the Average Linkage method, Complete Linkage method, Centroid method, McQuitty method, Median method, Single Linkage method, and Ward's method.

By selecting these dissimilarity measures to perform cluster analysis, we obtain a distance matrix from the inter-individual dissimilarity, which illustrates the degree of similarity between the classified data. From the inter-cluster dissimilarity, we obtain (i) a cluster formation history that indicates the sequential order of grouping all data into a hierarchical structure, and (ii) the distances between each cluster. These results can be collectively presented in the form of a dendrogram.

For instance, Figure 1 is a dendogram based on the distribution of Table 2. The height scale tells us that Dutch and Polish form a cluster "earlier" than other objects to form a cluster with Serbian, then {German, Russian}, finally English.

As stated, there are several ways to measure inter-individual dissimilarity, and in this case, the Canberra distance method, which is often used in prior research [16], is adopted. Similarly, the Ward method, which is adopted in prior research [16], is used for the method of clustering.

In conducting hierarchical cluster analysis, multiscale bootstrap resampling, a computer-based simulation method [10], was used to calculate p-values [16]. There are two types of p-values, the AU (Approximately Unbiased) p-values (on the left, normally in red) and the BP (Bootstrap Probability) value (on the right, normally in green). Clusters with high p-values are strongly backed by the data.

3.2.2. Correspondence analysis

In this study, we use correspondence analysis to identify patterns of association and disassociation in linguistic data. As explained by Glynn [20] in relation to the correspondence analysis of linguistic data, correspondence analysis is defined as "a multivariate exploratory space reduction technique for categorical data analysis" [20, p.443].

Correspondence analysis enables analysts to visualize the relationships between categories. By mapping the relative positions of these categories in two-dimensional or three-dimensional space, one can discern which categories are proximate to one another and which are distant. As a starting point, Glynn provides a sample in Table 3 [20, p.453]. One should prepare frequency tables similar to Table 4 [20, p.453]. Using Table 3 as an example, one can create a cross-tabulation of the second column labeled "Verb" with the third column labeled "Gram. Category" [20, p.453]. To bring out the main patterns or structures in the data, CA projects the information from this cross-tabulation into a lower-dimensional space. This dimension reduction is achieved through specific mathematical techniques, such as singular value decomposition, and is further visualized by plotting these positions on a two-dimensional or three-dimensional graph [20, p.456]. Within this projected space, categories plot-

Table 3 An example of categorical data [20, p.453]

Example	Verb	Gram. category	Person	Ind. obj. semantics
example1	think	Perfective	1st	Human
example2	suppose	Modal	3rd	Concrete_Thing
example3	suppose	Perfective	3rd	Abstract_State_of_Affairs
example4	believe	Imperfective	1st	Concrete_Activity
example5	say	Imperfective	3rd	Abstract_State_of_Affairs
example6	talk	Modal	1st	Concrete_Thing
example7	suppose	Imperfective	1st	Concrete_Activity
example8	speak	Perfective	1st	Human
to 575 examples	...	...	...	...

Table 4 An example of a numerical cross-tabulation contingency table [20, p.453]

	believe	think	suppose	say	speak	talk
Perfective	32	28	22	16	20	14
Imperfective	24	24	34	42	49	44
Modal	44	52	48	29	26	27

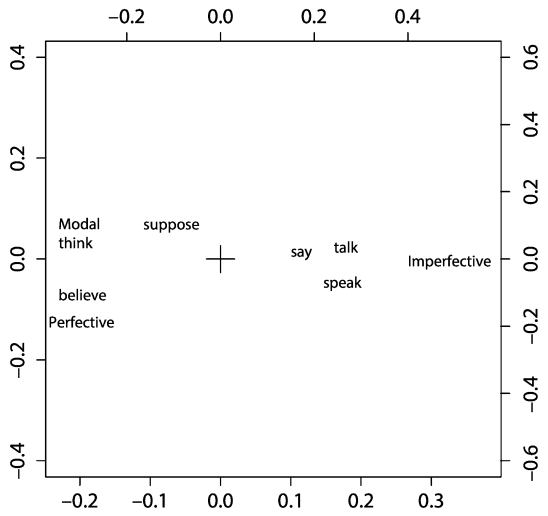


Figure 2 An example of correspondence analysis [20, p.456]

ted closely together exhibit similar profiles, while those plotted at greater distances display contrasting profiles.

In this study, we also used the frequency table that was used in the cluster analysis to conduct correspondence analysis. Combining cluster analysis with correspondence analysis is effective to confirm how each variable contribute to forming a cluster. Though cluster analysis shows how close each objects are, those clusters do not tell us how each of them was formed. As explained, correspondence analysis can visualise the closeness of each variable in a form of scatter plot (Figure 2), which can visualise the variables that contribute to forming clusters of each subarray.

3.2.3. Logistic regression analysis

We employed logistic regression analysis to explore the relationship between the realisation of ⟨EXPLAINER⟩ and document types. Logistic regression analysis is a kind of regression analysis whose response variable is categorical (i.e., every sentence involves co-composition or not) [17, 36]. A family of regression analysis can reveal the difference in the data and predict the kind of variables contributing to the distribution of response variables. Logistic regression analysis aims to obtain the formula in the form of (7), where p is the probability of the response variable, e is Napier's constant, α is an intercept, and β is the slope.

(7) $logit(x) = log_e(\frac{p}{1-p}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i$

Since quantitative analyses of language mainly deal with categorical variables, it

is possible to conduct a similar analysis by employing chi-square tests. However, statistical tests can only reveal correlations among variables, and analysts cannot always draw a prediction from the data. Logistic regression analysis can instead predict the kind of variables (i.e., β_i) that contribute to the distribution of the response variable (i.e., α).

3.3. Procedure

Firstly, we obtained all objects of the verb *explain* were extracted from BNC (tagged by CLAWS) using Sketch Engine [22]. The query yielded 6,570 attested cases. We randomly sampled 10% cases, which resulted in 657 cases in total. Secondly, we annotated all 657 cases manually to investigate how `Explaining` is realised by annotating frame semantic information defined in (3–4). Thirdly, we annotated other grammatical features in 657 cases. Since the genre is pre-set in the corpus, we did not have to annotate anything. Using R [35], we analysed the obtained data using the techniques introduced above.

4. Result

This section is structured as follows: §4.1 presents a quantitative evaluation of the results. As will be explained in detail, the analysis of linguistic features and frame semantic annotations yielded two distinct cluster solutions, which accords with our proceeding assumptions. In §4.2, logistic regression analysis was used to link each cluster with different register types.

4.1. Quantitative Results

In this section, we report the quantitative evaluations of obtained results. Tables 6 and 7 provide a summary of the values obtained through manual annotation. We utilized genres that had previously been established in the BNC corpus. For concision, the genre names have been abbreviated as listed in Table 5. In the following, we provide a joint result of cluster analysis and correspondece analysis in §4.1.1, and a result of confirmation using logistic regression analysis in §4.1.2

4.1.1. Cluster analysis & correspondence analysis

This subsection presents the joint results of the cluster analysis and correspondence analysis.

Cluster analysis is a method that groups data samples based on their similarity or dissimilarity, as measured by distance metrics [8, 9]. To perform cluster analysis, a relativised frequency table is necessary. In this study, Table 8 presents a cross-tabulation of genre distributions, while Table 9 displays grammatical features. The Ward method was employed as the clustering algorithm, and the Canberra distance

Table 5 Abbreviations of genre names

Original names	Abbreviated names
Applied science	AS
Arts	A
Belief & thought	B&T
Commerce & finance	C&F
Imaginative	I
Leisure	L
Natural & pure science	N&PS
Social science	SS
Unknown	U
World affairs	WA

Table 6 Cross tabulation of subarrays and grammatical features

	Freq	Act	Pas	Dec	Imp	Ass	Quo	wEx	w/oEx
En + Ed	210	156	54	210	0	210	0	210	0
Exp + Top	191	187	4	191	0	139	52	191	0
Top	97	75	22	83	14	97	0	62	35
Exp + Aud + Top	72	68	4	72	0	68	4	72	0
Top + Med	43	34	9	42	1	42	1	33	10
Ed	16	8	8	16	0	16	0	16	0
Exp + Top + Med	14	14	0	14	0	12	2	14	0
Aud + Top	12	5	7	12	0	12	0	5	7
Exp + Aud + Top + Med	2	2	0	2	0	2	0	2	0

Table 7 Distribution of genres

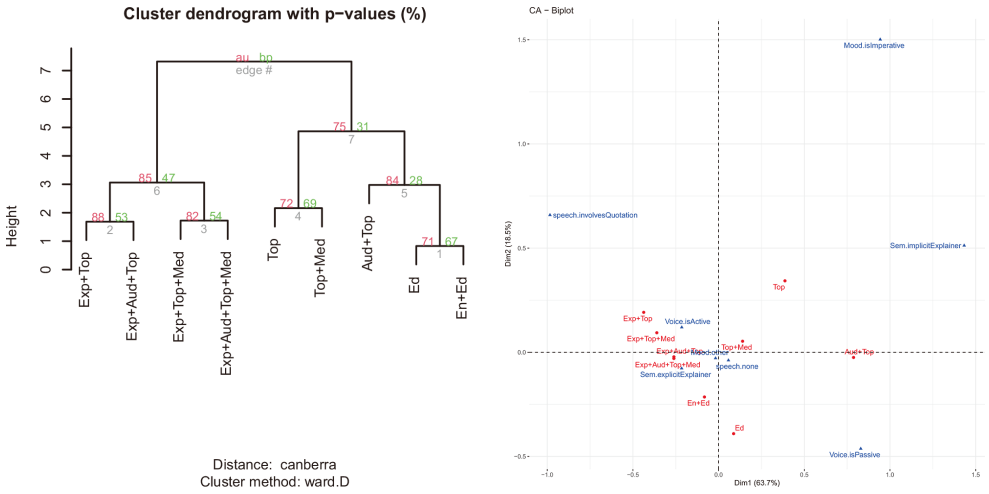
	AS	A	B&T	C&F	I	L	N&PS	SS	U	WA
En + Ed	26	17	18	17	14	5	41	45	2	25
Exp + Top	10	26	6	18	29	29	5	29	13	26
Top	8	5	10	18	4	6	4	27	0	15
Exp + Aud + Top	4	7	1	9	20	6	0	8	7	10
Top + Med	2	4	2	13	1	3	4	6	3	5
Ed	2	1	1	0	3	1	0	4	0	4
Exp + Top + Med	2	1	1	1	1	2	1	2	1	2
Aud + Top	2	0	0	2	1	1	1	2	0	3
Exp + Aud + Top + Med	0	0	0	0	0	1	0	1	0	0

Table 8 Distribution of genres (with standardised values)

	AS	A	B &T	C&F	I	L	N &PS	SS	U	WA
Ed	0.125	0.063	0.063	0.000	0.188	0.063	0.000	0.250	0.000	0.250
En+Ed	0.123	0.081	0.085	0.081	0.066	0.024	0.194	0.218	0.009	0.118
Top	0.082	0.052	0.103	0.186	0.041	0.062	0.041	0.278	0.000	0.155
Top+Med	0.047	0.093	0.047	0.302	0.023	0.070	0.093	0.140	0.070	0.116
Aud+Top	0.167	0.000	0.000	0.167	0.083	0.083	0.083	0.167	0.000	0.250
Exp+Top	0.053	0.137	0.032	0.095	0.153	0.153	0.026	0.147	0.068	0.137
Exp+Top+Med	0.143	0.071	0.071	0.071	0.071	0.143	0.071	0.143	0.071	0.143
Exp+Aud+Top	0.056	0.097	0.014	0.125	0.278	0.083	0.000	0.111	0.097	0.139
Exp+Aud+Top+Med	0.000	0.000	0.000	0.000	0.000	0.500	0.000	0.500	0.000	0.000

Table 9 Cross tabulation of grammatical features (with standardised values)

	Voice		Modality		Quotation		⟨EXPLAINER⟩	
	Passive	Active	Imperative	Declarative	Quotative	Assertive	without	with
Ed	0.500	0.500	0.000	1.000	0.000	1.000	0.000	1.000
En+Ed	0.256	0.744	0.000	1.000	0.000	1.000	0.000	1.000
Top	0.227	0.773	0.144	0.856	0.000	1.000	0.361	0.639
Top+Med	0.209	0.791	0.023	0.977	0.023	0.977	0.233	0.767
Aud+Top	0.583	0.417	0.000	1.000	0.000	1.000	0.583	0.417
Exp+Top	0.021	0.979	0.000	1.000	0.274	0.726	0.000	1.000
Exp+Top+Med	0.000	1.000	0.000	1.000	0.143	0.857	0.000	1.000
Exp+Aud+Top	0.056	0.944	0.000	1.000	0.056	0.944	0.000	1.000
Exp+Aud+Top+Med	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000

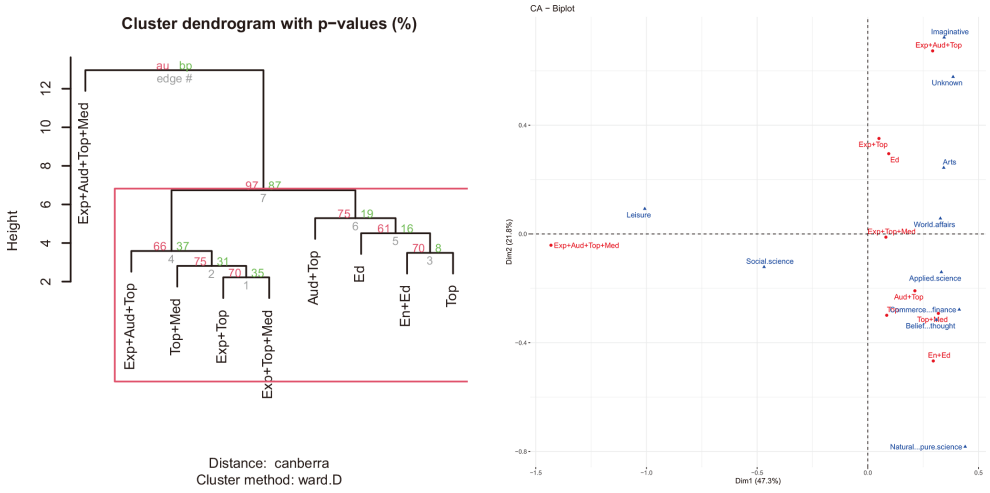


(a) Cluster dendrogram (b) Correnspondece analysis

Figure 3 Analysis of grammatical variables

was used to measure distances. Furthermore, the bootstrap method was utilised to validate the dendrograms [10]. For each condition, we present the result of cluster analysis with that of correspondence analysis.

Analysis of grammatical variables Figure 3a depicts the cluster dendrogram based on grammatical variables. The dendrogram displays the formation of a group including ⟨EXPLAINER⟩ on the left-hand side, while on the right-hand side, groups excluding ⟨EXPLAINER⟩, specifically ⟨EXPLANAN⟩ and ⟨EXPLANANDUM⟩, are formed. Though the overall structure is quite intuitive, their grouping was not supported by the bootstrap measures. The bi-plot of correspondence analysis (shown as Figure 3b) presents the results of the correspondence analysis, which reveal that VOICEISACTIVE and EXPLICITEXPLAINER correspond to the ⟨EXPLAINER⟩ group on the left-hand side of the graph. The non-⟨EXPLAINER⟩ groups (AUD+TOP, TOP, TOP+MED) are situ-



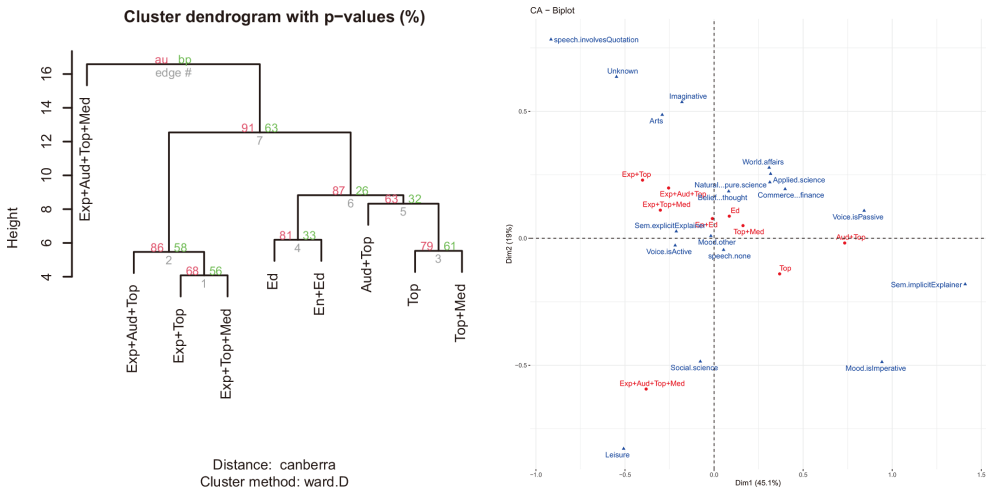
(a) Cluster dendrogram (b) Correnspondece analysis

Figure 4 Analysis of genre variables

ated on the right-hand side. Meanwhile, the ⟨Explanan⟩+⟨Explanandum⟩ groups are located on the lower part of the graph. This result suggests that each grammatical feature contribute to realising ⟨Explainer⟩.

Analysis of genre variables Figure 4a is the cluster dendrogram utilising only genre variables. The dendrogram reveals a tendency for ⟨Explainer⟩ groups to form on the left-hand side and non-⟨Explainer⟩ groups to form on the right-hand side. As the figure shows, except for "Ed+Aud+Top+Med", other subarray formed a significant group. The bi-plot of correspondence analysis (shown as Figure 4b) depicts that Exp+Aud+Top+Med corresponds to Leisure and Social Science. On the upper right-hand side of the graph, non-academic genres are grouped and correspond to the ⟨Explainer⟩ group. Meanwhile, academic genres are situated on the lower right-hand side and correspond to the non-⟨Explainer⟩ group.

Analysis of all variables Figure 5a is the cluster dendrogram utilising all variables (i.e., grammatical features and genres). This dendrogram is divided into two groups, with and without ⟨Explainer⟩, a pattern that can also be observed in other dendrograms. Like the the dendgram based on grammatical features, no significant clusters were observed. The bi-plot of correspondence analysis (shown as Figure 5b) suggest that involveQuotation corresponds to Unknown, Imaginative, and Arts on the upper left-hand side of the graph. On the lower left-hand side, Exp+Aud+Top+Med corresponds to Social Science and Leisure. VoiceIsPassive, implicitExplainer, and MoodIsImperative correspond to the non-



(a) Cluster dendrogram (b) Correnspondece analysis

Figure 5 Analysis of all variables

⟨EXPLAINER⟩ group (AUD+TOP, TOP, TOP+MED) on the right-hand side. Meanwhile, VOICEISACTIVE, EXPLICITEXPLAINER, and MOOD OTHER correspond to the ⟨EXPLAINER⟩ group on the left-hand side. Academic genres are situated on the upper right-hand side of the graph and do not correspond to the ⟨EXPLAINER⟩ group. Finally, ED and EN+ED are situated at the center of the graph.

4.1.2. Logistic regression analysis

This section reports the result of the logistic regression analysis. Logistic regression analysis can reveal the categorical response variable and various predictors. In addition to automatic classification of *explain*, this study also explored the relationship between the presence of ⟨EXPLAINER⟩ and the academicity of genres. The statistical analysis of these variables revealed that ⟨EXPLAINER⟩ is more likely to be present in academic genres (54.5%) than in non-academic genres (30.7%).

As discussed in §2.2.1, thorough annotation of linguistic texts reveals many aspects of language usage. Cluster analyses suggested the correlation between the presence of ⟨EXPLAINER⟩ and the academicity of genres. These values can be summarised as Table 10, which is visualised as a mosaic plot (the left panel) in Figure 6. As the mosaicplot shows, the proportions of ⟨EXPLAINER⟩'s presence correlate with those of academic genres. To confirm this tendency statistically, **logistic regression analysis** was carried out [36].

Logistic regression analysis was carried out using R [35]. The presence of ⟨EXPLAINER⟩ and the academicity of genres are categorical (binary). The result of the analysis is summarised as Table 11, which indicates that the model is statis-

Table 10 Cross tabulation of genre and presence of ⟨Explainer⟩

	Non-academic	Academic
⟨Explainer⟩ is absent	146	233
⟨Explainer⟩ is present	175	103

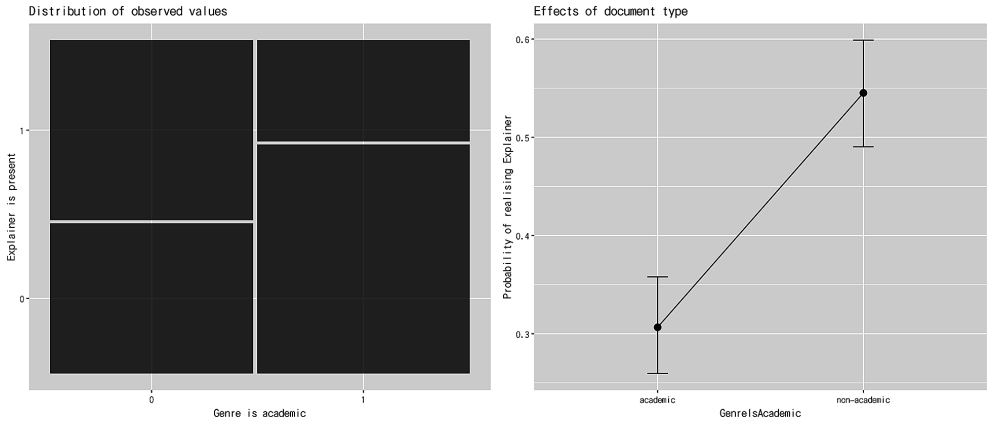


Figure 6 Mosaic plot of Table 10, and effect plot of logistic regression analysis

Table 11 Coefficient table of ⟨Explainer⟩ is present and the academicity of genres

	Estimate	Std. Error	z value	$\Pr(>\|z\|)$	p-value
(Intercept)	0.181	0.112	1.616	0.106	
genreIsAcademic1	−0.998	0.163	−6.12	$9.35e-10$	***

tically significant by incorporating the categorical variable (i.e., academic genres vs non-academic genres). We visualised the effect of genre type (i.e., academic vs. non-academic) and response variable (i.e., ⟨Explainer⟩ is realised or not) as the part of Figure 6 (the right panel) with confidence intervals.

The formula in (8) can be used to compute the probability of each response variable. (8a) corresponds to the probability of present ⟨Explainer⟩ in the academic genres, and (8a) to the probability of present ⟨Explainer⟩ in the non-academic genres. These values are plotted respectively on the right panel of Figure 6.

$$(8)\quad p = \frac{1}{1 + e^{-(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_i x_i)}}$$

$$\text{a.}\quad p = \frac{1}{1 + e^{-(0.182-0.9975*1)}} = 0.547$$

$$\text{b.}\quad p = \frac{1}{1 + e^{-(0.182-0.9975*0)}} = 0.307$$

4.2. Qualitative evaluation of the result

This section explores some of the qualitative findings in our study, namely (i) dif-

ferentiation patterns of *explain* and (ii) the quotative use of *explain*. All examples from the corpus are followed by their file names in parenthesis, and [...] denotes an abbreviation for convenience.

The differentiation patterns of *explain*: As introduced in §3.1.1, we assumed communicative and scientific senses of *explain*. Examples in (9) are annotated with frame elements defined in (4). In these cases, the grammatical subjects of *explain* are qualitatively different. ⟨EXPLAINER⟩ is a sentient entity, while ⟨EXPLANAN⟩ is not. Since the communicative sense of *explain* clearly expresses the verbal interaction of entities (typically humans), the differentiation of these patterns is quite straightforward.

(9) a. Now he says he wrote to S & N chairman, Alec Rankin, asking him ⟨EXPLAINER⟩ to **explain** fully the reasons for switching away from diesel ⟨TOPIC⟩. (AKM 1236)

b. Several geological factors ⟨EXPLANAN⟩ may **explain** this apparent discrepancy ⟨EXPLANANDUM⟩. (CRM 10791)

The quotitative use of *explain*: In the communicative sense of *explain* (as defined in (3a)), some use of *explain* were synonymous with the verb *say*. For instance, *explain* in (10) takes quoted contents as its grammatical object, which is annotated as ⟨TOPIC⟩. Strangely, all instances of quotative use ($n = 52$) involved the inversion of grammatical objects.

(10) [...] "You must understand," ⟨TOPIC⟩ **explained** Mrs Puri ⟨EXPLAINER⟩, [...] (H89 88)

5. Discussion

This section discusses the theoretical implications of our result. §5.1 compares our results with previous studies and argues that detailed corpus analysis can reveal "deeper" aspects of linguistic usage. §5.2 further explores the implication of commitment to linguistic theory (namely frame semantics) by contrasting two classes of *explain*.

5.1. Comparison between our results and previous studies

This section expounds upon the methodological implications of our research by comparing it to previous work [31]. Specifically, there are differences between (i) the methods of data collection and (ii) the methods of data analysis.

First, while Overton [31] collected various instances with the verb *explain* from different sources, the balancedness of the collection can be questioned. In contrast, our research utilises a large and balanced corpus, the British National Corpus (BNC), which contains 6570 cases where the word 'explain' is used as a verb and randomly selects 657 of those cases. As a result, we suceeded in identifying an "ordinal" uses of *explain* in contrast to "academic" uses. This tendency was clear especially in the result of logistic regression analysis (§4.1.2).

Second, Overton [31] simply tallied the tokens of the words related to *explain*. While this approach can handle the quantitative side of analysis and is suitable for words with only one meaning, it may be insufficient for analysing polysemous concepts like *explain*. Our research, on the other hand, combines quantitative analysis with the method of cognitive semantics to identify the concept's polysemy and describe fine-grained linguistic features.

Our research endeavours attempted to naturalise concept analysis by integrating it with linguistic methods (e.g., corpus linguistics and cognitive semantics). Both Overton's [31] approach and our approach share the methodology of introducing text analysis to concept analysis from a naturalistic perspective. However, there is a difference in whether linguistic methodologies are employed or not. If the linguistic analysis is the only way to understand linguistic practices in philosophical studies, then Overton's naturalistic approach may be insufficient as it does not utilise linguistic methods.

5.2. Two subclasses of `Explaining`

This section discusses some of the assumptions and consequences of this study.

As suggested from the definition of $\texttt{Explaining}_{\{1,2\}}$, one of the assumptions of this study is a polysemous nature of the verb *explain*. Generally, a frame is characterised as a set of semantic roles and their relations among them. In other words, frames are descriptions of events in the world. $\texttt{Explaining}_1$ and $\texttt{Explaining}_2$ share the same parent frame, which works as a node in a network of frames. The network of frames can be equated with the ontology of events.

Our analysis revealed that natural language exhibits a subtle nature of distinguishing these two subtypes of *explain*. The combination of manual annotation and statistical techniques enabled the exploration of the relationship between language use and events. Exploring and interpreting the collocational tendency can reveal the overall semantic tendency of expressions. However, it can only scratch the surface of the complex nature of language use.

Though a finer-grained analysis was possible because of frame semantic annotations, the organisation of frames must be discussed thoroughly. To achieve a more precise analysis, ontological analysis of *explain* must be conducted [26, 27, 28]. Our analysis aimed to examine the nature of discourse involving the expression *explain*

while imposing an ontological structure. The proposed strategy for quantitative concept analysis offers a bridge between epistemological and ontological enquiry, which was possible because of the commitment to linguistic methodologies.

Dennett [6, p.3–4] emphasised that ontological and epistemological questions must be answered simultaneously. The richness in epistemology and ontology has a trade-off relationship. If we "know" too many things in the world, it becomes doubtful to assume our knowledge is equated with categories (or things) in the world. Likewise, if there are too many things in the world, how we "know" so many things becomes another problem. Our analysis revealed that how we "explain" things is inseparable from what "explain" is.

6. Conclusion

This study aimed to observe the conceptual essence of the verb *explain* by means of a blend of manual annotations and statistical analyses. Consequently, our findings suggest that two varieties of *explain* are distributed across different genres, signifying that conventional philosophers of science dealt only with a specific subtype of *explain*. Besides presenting a more nuanced analysis than the currently accessible ones, our approach provides a linkage between ontology and epistemology.

Several issues remain unsolved. Firstly, we need to explore the ontological nature of *explain*. This study employed the frame structure with two child frames with two to four frame elements. However, we must ensure this structure is sound enough for other usages of *explain*. Secondly, this study did not explore the collocational tendency of each child frame. To compare our result with Overton's [31], we must conduct a thorough collocational analysis.

A. Data set of this study

All the data set and codes for statistical analysis are available on: `https://osf.io/znr7x/?view_only=5cfebf52ca4d4546b7594432b7051551`

References

[1] J. M. Byron, Whence philosophy of biology?, The British Journal for the Philosophy of Science, 58(3), 409–422, 2007.
[2] P. Carruthers, S. Stich, and M. Siegal, (eds.). The Cognitive Basis of Science. Cambridge University Press, 2002.
[3] L. Chartrand, Modeling and corpus methods in experimental philosophy. Philosophy Compass, e12837, 2022.
[4] A. D. Cruse, Lexical Semantics. Cambridge University Press, 1986.
[5] D. C. Dennett, Consciousness Explained. Penguin, 1993.

[6] D. C. Dennett, Kinds of Minds: Toward an Understanding of Consciousness. Basic Books, 1996.

[7] D. C. Dennett, From Bacteria to Bach and Back: The Evolution of Minds. Penguin, 1993.

[8] G. Desagulier, Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics. Springer, 2017.

[9] D. Divjak, and N. Fieller. Cluster analysis: Finding structure in linguistic data. In D. Glynn, and K. Fischer (eds.), Corpus Methods for Semantics: Quantitative Studies and Polysemy and Synonymy, 405–441, Mouton de Gruyter, 2014.

[10] J. Egbert, and L. Plonsky, Bootstrapping techniques. In M. Paquot, and S.Th. Gries (eds.), A Practical Handbook of Corpus Linguistics, 593–610, Springer, 2020.

[11] C. J. Fillmore, "Corpus linguistics" vs. "computer-aided armchair linguistics". In J. Svartvik (ed.), Directions in Corpus Linguistics: Proceedings from a 1991 Nobel Symposium on Corpus Linguistics, 35–66, Mouton de Gruyter, 1992.

[12] C. J. Fillmore, C. Wooters, and C. F. Baker, Building a large lexical databank which provides deep semantics. In Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation, 3–26, 2001.

[13] C. J. Fillmore, C. R. Johnson, and M. R. L. Petruck, Background to FrameNet. International Journal of Lexicography, 16(3), 235–250, 2003.

[14] C. J. Fillmore, and C. F. Baker, A frames approach to semantic analysis. In B. Hein, and H. Narrog (eds.), The Oxford Handbook of Linguistic Analysis, 791–816, Oxford University Press, 2012.

[15] S. Th. Gries, Corpus-based methods and cognitive semantics: The many senses of *to run*. In S.T. Gries, and Anatol Stefanowitsch (eds.), Corpora in Cognitive Linguistics: Corpus-Based Approach to Syntax and Lexis Approaches to Syntax and Lexis, 57–99, Mouton de Gruyter, 2006.

[16] S. Th. Gries, Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics. The Mental Lexicon 5(3), 323–346, 2010.

[17] S. Th. Gries, Statistics for Linguists with R: A Practical Introduction. De Gruyter Mouton, 2021.

[18] S. Th. Gries, Quantitative corpus methods of cognitive semantics. In T. Li (ed.), Handbook of Cognitive Semantics, 328–350, Brill, 2023.

[19] D. Glynn, The many uses of run: Corpus methods and socio-cognitive semantics. In D. Glynn and J. A. Robinson (eds.), Corpus Methods for Semantics, 117–144. John Benjamins, 2014.

[20] D. Glynn, Correspondence analysis: Exploring data and identifying patterns. In D. Glynn, and K. Fischer (eds.), Corpus Methods for Semantics: Quantitative Studies and Polysemy and Synonymy, 443–485, Mouton de Gruyter, 2014.

[21] L. A. Janda (ed.), Cognitive Linguistics: The Quantitative Turn, Mouton de Gruyter, 2013.

[22] A. Kilgarriff, and V. Baisa, J. Bušta, M. Jakubíček, V. Kovvář, J. Michelfeit, P. Rychlý, and V. Suchomel, The Sketch Engine: Ten Years On. Lexicography, 1(1), 7–36, 2014.

[23] E. Machery, Experimental philosophy of science. A companion to experimental phi-

losophy, 475–490, 2016.
[24] E. Machery, and K. Cohen, An evidence-based study of the evolutionary behavioral sciences. The British journal for the philosophy of science, 63(1), 177–226, 2012.
[25] T. McEnery, and A. Hardie, Corpus Linguistics: Method, Theory and Practice. Cambridge University Press, 2011.
[26] R. Mizoguchi, Tutorial on ontological engineering Part 1: Introduction to ontological engineering. New Generation Computing, 21(4), 365–384, 2003.
[27] R. Mizoguchi, Tutorial on ontological engineering Part 2: Ontology development, tools and languages. New Generation Computing, 22(1), 61–96, 2004.
[28] R. Mizoguchi, Tutorial on ontological engineering Part 3: Advanced course of ontological engineering. New Generation Computing, 22(2), 193–220, 2004.
[29] F. Moretti, Distant reading. Verso Books, 2013.
[30] M. L. Murphy, Lexical Meaning. Cambridge University Press, 2010.
[31] J. A. Overton, "Explain" in scientific discourse. Synthese, 190(8), 1383–1405, 2013.
[32] C. H. Pence and G. Ramsey, How to do digital philosophy of science, Philosophy of Science, 85(5), 930–941, 2018.
[33] W. V. Quine, Epistemology naturalized. In Ontological Relativity and Other Essays, 69–90. Columbia University Press, 1969.
[34] W. V. Quine, Pursuit of Truth, Revised edition, Harvard University Press, 1992.
[35] R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2022. URL: `https://www.R-project.org/`.
[36] D. Speelman, Logistic regression: A confirmatory technique for comparisons in corpus linguistics. In D. Glynn, and J. A. Robinson (eds.), Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy, 487–533. John Benjamins, 2014.
[37] P. Thagard, Computational Philosophy of Science. The MIT Press, 1987.
[38] P. Thagard, The Cognitive Science of Science: Explanation, Discovery, and Conceptual Change. The MIT Press, 2014.
[39] R. Ventura, Quantitative methods in philosophy of language. Philosophy Compass, 14(7), e12609, 2019.
[40] S. B. Weingart, Finding the history and philosophy of science. Erkenntnis, 80(1), 201–213, 2015.
[41] K. B. Wray, Philosophy of science: What are the key journals in the field?, Erkenntnis, 72(3), 423–430, 2010.

(Received 2022.12.31; Revised 2023.7.29; Accepted 2023.8.6)