Philosophy of Data Science for Corpus Linguistics:

A Pragmatistic Point of View

Kazuho KAMBARA* and Tsukasa YAMANAKA**

Abstract

This paper demonstrates pragmatic constraints involved in corpus linguistic studies. Both philosophers and linguists have long recognised the difficulties in characterising meaning. Despite the widely acknowledged difficulty, quantitative semantic analysis has been attempted. In this paper, how corpus linguists make pragmatic decisions is explained by introducing the degrees of specificity (i.e., is-a relation, or class inheritance) and granularity (i.e., part-whole relation, or mereological relation) in identifying and describing linguistically expressed concepts. We show that pragmatic constraints are ubiquitous in many aspects of quantitative semantic analyses.

Key words: Corpus Linguistics, Quantitative Corpus Method (QCM), Pragmatism, Collocation Analysis, Feature Analysis, Ontology, Naturalism, Specificity and Granularity

1. Introduction

This paper aims to show that pragmatic constraints are ubiquitous in the quantitative analysis of language, which suggests the Platonistic view of meaning does not fit into

* Ritsumeikan University
E-mail: kazy0324@pep-rg.jp
** Ritsumeikan University
E-mail: tsukasayamanaka@pep-rg.jp
Portions of this study was reported at Mebius Association of Language Studies at Kyoto University of Foreign Studies. We thank the audience there. We also thank the anonymous reviewers for their helpful feedbacks. This work has been supported by R-GIRO. The first author would like to thank Professor Stefan Th. Gries for sharing his data to partially replicate his analysis.

the research practice of quantitative linguistics. This section is structured as follows. Section 1.1 briefly overviews the background of pragmatism, and 1.2 introduces the goal and structure of this paper.

1.1. Scientificity and pragmatism

This paper deals with the connection between scientificity and pragmatism. Some may argue for the anti-scientificity of pragmatism. Historically, advocates of neo-pragmatism [57] regard pragmatism as "something that obstructs the progress of science", or more simply "anti-science". On the other hand, the original conception of pragmatism is inseparable from that of scientific methods [50, 51, 52, 56]. Some followers of pragmatism [30, 54] see clear connections with scientific enquiries. The confusing conflicts of scientificity and pragmatism probably come from the multi-faceted nature of pragmatism [3].

One of the central ideas in pragmatism is the limited nature of human beings. As a finite being, we do not have access to truly "objective" world (or, truth), therefore we exploit whatever available to understand and cope with it. Though scientific enquiries seem to provide objective "truth" about the world, pragmatic natures of our enquiry are inevitable in any field. This empirical nature of pragmatism concords with the bottom-up nature of quantitative corpus linguistics.

Since the goal of philosophy of science is to understand how scientists arrive at agreement (or, understandings of their target phenomena), revealing how pragmatic constraints in scientific enterprises is vital. Especially in quantitative semantic analysis, analysts must empirically come to agreement what the meaning of a given expression is. If we were able to access the truly objective, unquestionable meaning (i.e., Platonistic meaning), we would not have any problem in identifying and delimiting the meanings of given expressions. Pragmatism denies the existence and understandings of such meanings, since we do not have any means to access them [46]. We see that the elusive nature of symbols (including language) can also be the source of the ambivalence of pragmatism.

1.2. The goal and structure of this paper

Linguistic meaning has drawn much academic attention due to its "ungraspability". Traditionally, many philosophers and linguists, under the influence of formal logic, have attempted to analyse it by utilising mathematical tools. On the other hand, the widespread of computers provided linguists with easy access to analyse their data quantitatively. Linguists have shown that the use of corpora, collections of authentic language use, has effectively overcome some limitations of the traditional methods. In contrast to formal approaches to linguistic semantics, the rationale behind such

analyses has rarely been described[1]. We see that pragmatic constraints on analytical choices are pervasive in each stage of empirical semantic research and such choices support the pragmatistic conception of meaning. Moreover, we further explore the implications of pragmatic constraints.

In this paper, we limit ourselves to discussing common methodological procedures and choices in corpus linguistics. For this reason, we will not discuss statistical techniques often employed in corpus linguistic research papers.

This paper is structured as follows. Section 2 reviews and explains the quantitative analysis of linguistic meaning in recent years in contrast to traditional approaches in linguistics. Section 3 introduces the two degrees of precision in concept structuring, which will be used to describe the target selection process in corpus linguistic enquiries. Section 4 discusses how pragmatic constraints are involved in the quantitative analysis of meaning. Section 5 summarises and discusses possible future developments.

2. Introspection and quantitative corpus method

This section explains the methodological backgrounds in linguistics and introduces the common approaches in quantitative analysis. Section 2.1 reviews the conventional procedures in linguistics that motivated the quantitative strategy, and Section 2.2 some of the benefits and techniques of quantitative language analysis.

2.1. Introspective meaning

This section reviews and critically assesses assumptions of traditional frameworks to analyse linguistic meaning. Section 2.1.1 introduces the conventional methodology based on an analyst's intuitive judgement to minimal pairs. Section 2.1.2 points out the limitations of the traditional methods.

2.1.1. Identification of meaning

This section reviews the traditional method of identifying an expression's meaning, which relies on an analyst's introspective judgement.

Linguistics is characterised as a science of language, which means that principled methodologies are devised to adequately describe every aspect of language (namely, phonological, morphological, syntactical, semantical, and pragmatical aspects). To

[1] The work of McEnery and Brezina [42] is a notable exception. They attempted to reveal how Popperian conception of science accords with methodologies in corpus linguistics. As will be clear in the following sections, we emphasises various pragmatic constraints in research practices. Though pursuing the difference between their approach and ours would be fruitful, we do not discuss this matter any further for reason of space.

observe semantic features in language use, linguists traditionally performed **acceptability judgement** to **minimal pairs**. Acceptability judgement refers to intuitive judgement, whether the expression in question is acceptable as a natural language, and minimal pairs to two or more linguistic expressions that differ in one aspect.

For instance, when a linguist decides to reveal the characteristics of three synonymous words, {*funny*, *peculiar*, *comical*}, she makes up sentences like (1–2) to contrast each word to observe how "naturalness" differs from one another[2]. Since simple contemplation of lexical meaning can lead to naive and unreplicable conclusions, acceptability judgement can be seen as an objectified methodology to identify the meaning of expressions. From the data in (1–2), one can infer that *funny* and *peculiar* refer to abnormal states while *funny* and *comical* to silly, laughable states.

(1) a. My tummy feels a bit **funny** whenever I eat fish.
b. My tummy feels a bit **peculiar** whenever I eat fish.
c. ?? My tummy feels a bit **comical** whenever I eat fish.

[48, p.110]

(2) a. Anna told a hilariously **funny** joke.
b. ?? Anna told a hilariously **peculiar** joke.
c. Anna told a hilariously **comical** joke.

[48, p.110]

The utilisation of unacceptable (or ungrammatical[3]) sentences has been one of the most significant advances in linguistic science. This methodology can also be applied in the studies of other areas (e.g., phonology, morphology, syntax, and pragmatics). Independent of the kind of linguists' framework, acceptability judgement plays a vital role in observing the target expressions. Since such judgement is based on analysts' intuition, some linguist goes so far as to characterise linguistic semantics as a branch of phenomenology [59, p.4].

[2] One may argue that they are not synonyms since it is impossible to replace one another in every context. In a (classical) Chomskyan tradition [48, pp.110–113], synonyms are categorised into two groups: perfect synonyms and sense synonyms. The former corresponds to the replaceable pair of words in every context, and the latter to the replaceable pair of words in some contexts. Strictly speaking, *funny*, *peculiar*, and *comical* are sense synonyms. We ignore these finer-grained classification of synonymy for simplicity.

[3] In (classical) Chomskyan tradition [6], an ungrammatical sentence and an unacceptable sentence are treated differently. The former refers to an impossible array of expressions. In contrast, the latter simply refers to an undesirable array of words influenced by third factors such as conventions, morality, etc. For the sake of simplicity, we leave this distinction out of consideration.

2.1.2. Imagination as a limitation

This section points out some of the well-known limitations of naive acceptability judgements. Despite the seemingly promising outlook of the methodology, two limitations are well-recognised in the utilisation of acceptability judgement. The first is the problem of generality, and the second is that of imagination as a constraint. We argue that the second problem is more severe than the other and suggest a possible solution.

The analytical procedure sketched in Section 2.1.1 presupposed the homogeneity of acceptability judgement between speakers. Arguments based on the introspective judgement rely heavily on analysts' intuition, which can sometimes differ from non-experts' judgement. For instance, a resultative construction, as exemplified in (3), refers to the movement of the napkin caused by the actor's sneeze. This example is well-known in the linguistics circle for its theoretical importance. However, it has been known that some non-experts tend to judge (3) as ungrammatical for the implausibility of the expressed situation. The generality of acceptability judgement is not as stable as linguists wished it to be [19, p.67]. However, the limitation can (relatively easily) be overcome by adopting psychologically plausible experimental procedures.

(3) He sneezed the napkin off the table.

[31, p.55]

Even after resolving the issue of generality by employing experimental procedures, how one creates appropriate sentences remains one of the challenges in the traditional analytical framework. As suggested from the discussion in Section 2.1.1, preparing the appropriate sentences (or phrases) to observe the characteristics of the target expression is the critical factor for a linguist to conduct sound research[4]. Despite its importance, the procedure of creating sentences is not as straightforward as one may assume.

The most severe issue in creating sentences is that analysts' imagination determines the range of observable data. Since it is difficult to implement the objective procedure to create the appropriate expressions, an analyst has no choice but to use her imagination to create her data set. However, if a scientist can fabricate her data freely, the plausibility of the intellectual enquiry can be questioned. Some linguists have a talent for coming up with the "right" sentences, while some do not.

To solidify the ground for observation, most linguists make use of **corpus**. Corpus (or corpora [plural]) is a machine-readable collection of authentic language usage. Its size can vary from one corpus to another. For instance, English Web 2020

[4] For instance, a linguist such as J. D. McCawley is well known for his thorough assessment of a wide range of examples [39, 40].

(EnTenTen20) contains 36,561,273,153 tokens of words [32], while one of the most popular corpora in linguistics circle, British National Corpus (BNC) "only" contains 96,052,598 tokens[5]. Corpora's contents differ because one has to decide the kind of data to incorporate into the building process.

Many scholars have found the use of corpora in linguistic research very effective because it allows one to observe a wide range of data. Fillmore emphasises that the corpus data forces linguists to recognise things that they are unlikely to have noticed otherwise [11, p.45]. Moreover, the utilisation of corpora in linguistic research enabled quantitative analysis.

Linguists who heavily rely on their intuitions are sometimes called armchair linguists. Though it is impossible to eliminate the use of introspection from the linguistic research program, how and when such judgement should be conducted must be discussed carefully [27, p.337].

2.2. Quantitative corpus methods (QCM)

This section reviews two major approaches in the quantitative corpus methods, namely: (i) collocational analysis, and (ii) feature analysis.

The quantitative corpus methods (QCM for short) are methodologies using corpora to analyse linguistic data quantitatively [26]. As discussed in Section 2.1.2, the traditional approach to linguistic meaning was conducted by relying heavily on the introspective judgement of an analyst. One of the challenges posed by this approach is that the analysts' imagination determines the range of observable data. The dissatisfaction with the traditional approach accelerated interest in quantitative methods in (especially cognitive) linguistics [33][6]. For this reason, corpus linguistics has drawn much academic attention[7].

As stated, corpus linguistics is often characterised as a methodology to analyse language employing a large-scale data set. Conducting a linguistic analysis with a large-scale corpus allows an analyst to encounter various examples. For instance, the verb *run* is notorious for having many senses [20, 29]. A quick query into corpus data[8] allows an analyst to observe various uses of *run* in different contexts. The verb *run* in (4a) refers to the fast pedestrian motion by a person, (4b) to the fast pedestrian motion by an animal [a horse], (4c) to the continuation of an event, and (4d) to the

[5] These numbers can change depending on the kind of segmentation algorithms since word segmentation is not always a straightforward task. See discussion in Gries [25, pp.12–14]

[6] Janda reports an increase of quantitative papers in the journal Cognitive Linguistics [33, p.13].

[7] For the reason of the space, we do not offer a detailed characterisation of corpus linguistics. For a more consistent coverage of this subject, refer to [41, 58].

[8] (4) is a group of sentences including the verb *run* from BNC (Tagged by CLAWS). Codes with parentheses indicate the source file in BNC data set.

Table 1 Observational differences in collocation and feature analysis of corpora [28, p.311]

	Collocation	Feature
Stage 1: Analysis of data	Objective	Subjective
Stage 2: Interpretation of analysis	Subjective	Objective

management of an event. Obtaining a wide range of data provides an opportunity to observe unbiasedly.

(4) a. I **ran** back and turned him, ever so gently, to get at his hands. (CS4 1297)
b. [...] but the fact that he **ran** so well last time at Aintree has kept him short in the market. (HJ3 1639)
c. [...] The exhibition, **running** until 27 June, [...] (CKW 823)
d. I have been **running** expeditions over the last 29 years [...] (K5A 2471)

Corpus linguistics has developed many effective methodologies to analyse language usage quantitatively. In linguistics, a corpus is used to reveal the characteristics of the target expression. The selection of target expression is often motivated by theoretical predictions. To conduct a quantitative corpus linguistic research, one has to decide on one of the two approaches (which sometimes can be mixed), namely **collocation analysis** and **feature analysis**. These approaches are defined in (5).

(5) a. **Collocation Analysis:** Quantitative evaluation of the target expression in relation to the expressions occurring around the target.
b. **Feature Analysis:** Quantitative evaluation of the target expression in relation to the result of (manual or automatic) annotations.

These two approaches are complementary to each other and have both strengths and limitations. Glynn [28, p.311] summarised the difference between these two approaches as Table 1. Characteristics of each approach are described in the following sections to demonstrate the strength of QCM. We review collocation analysis in Section 2.2.1 and feature analysis in Section 2.2.2.

2.2.1. Collocation analysis

This section explains how analysts carry out collocation analysis. As described above, in collocation analysis, analysts observe the kind of expressions that occurs around the target. Though it is tricky to determine the type of collocational strength appropriate for the analysis [24], an expression's collocations can be seen as good indicators of meaning. Though this analysis can be carried out objectively, its interpretation can sometimes be challenging.

Collocation analysis is often tied to Firth's famous dictum that "[y]ou shall

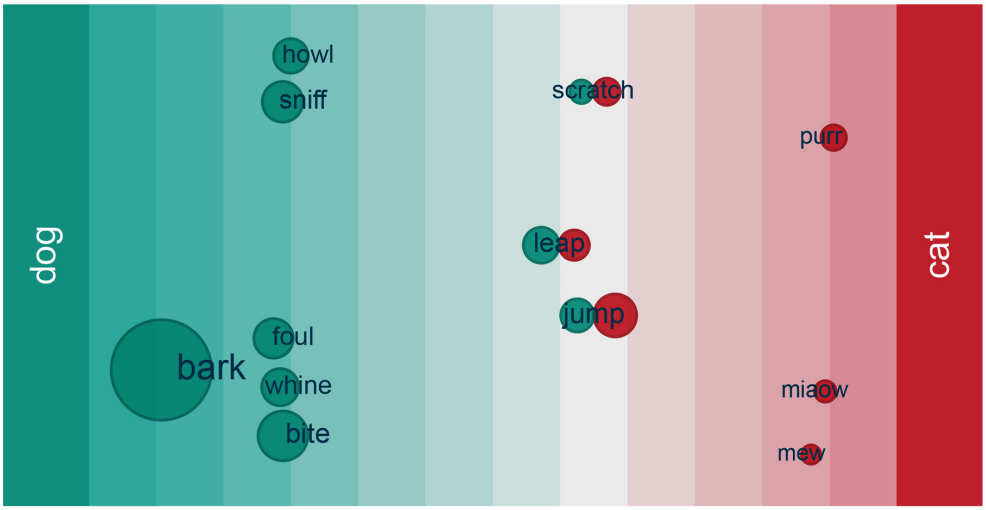

Figure 1 Collocational differences between *dog* and *cat*

know a word by the company it keeps" [16, p.11]. As discussed in Section 2.1, when an analyst attempts to reveal the semantic characteristic of an expression without consulting any database, it is challenging to distinguish inadequate inferences from adequate ones due to the lack of evidence. However, once an analyst observes the collocational information, she is likely to find semantically-motivated results.

Collocations tell analysts a lot about the target expression(s). When an analyst attempts to find the difference between *dog* and *cat*, collocational measures can be one of the most effective indicators of meaning. By using Sketch Engine [35], verbs with *dog* and/or *cat* as their subjects (in British National Corpus (BNC)) are visualised as Figure 1. In Figure 1, verbs that are likely to occur with *dog* and *cat* are shown in green and red circles, respectively. Each size of the circles corresponds to their raw frequencies in the corpus. Figure 1 shows us that *dog* is likely to occur with verbs like *bark* or *howl* while *cat* with *purr* or *miaow*. The collocations clearly tell us what *dog* or *cat* do in our world.

However, inferring common knowledge from collocations alone is a slippery slope. Take a pair of synonyms, *sofa* and *couch*. Since they are interchangeable in almost all contexts, as shown in (6), an analyst may infer that there are no collocational differences in this pair. However, this kind of conclusion is wrong in most cases.

(6) a. Alice bought a {sofa, couch}
b. Alice sold a {sofa, couch}
c. Alice destroyed a {sofa, couch}

Another look at data in English Web 2020 (EnTenTen20) tells us that though

sofa and *couch* share many collocations (e.g., *perfect, beautiful, warm, empty*), not all words are shared between *sofa* and *couch*[9]. The resulting data shows that adjectives like *ugly* and *green* only modify *couch* while adjectives like *handmade* and *lovely* only modify *sofa*. Though it is interesting to see such a skewed result, it is hasty to draw an ontological conclusion like "things we call *sofa* (or *couch*) can only be *handmade* or *lovely* (or *ugly* or *green*)".

Frequent collocations are good indicators of the target's meaning, but such results cannot always be equated with the target's necessary or sufficient conditions. Learning common knowledge from collocations has been attempted in natural language processing by (somewhat) "normalising" distributional features. However, naive collocation analyses only reveal "how the expression is used" rather than "what the expression means". Many studies have pointed out that semantically similar expressions are also distributionally similar [22, 36], which holds some truth. However, it is still difficult to distinguish the collocations that significantly reflect the target's meaning from those that do not.

2.2.2. Feature analysis

This section explains how analysts carry out feature analysis. As explained above, in feature analysis, analysts annotate expressions with meta-expressions. Annotations can be done manually or automatically, depending on the nature of meta-expressions. The precision between these two annotations is still debatable. More importantly, designing a consistent annotation framework is challenging in semantics (and in pragmatics). Despite such challenges, the evaluation of meta-expressions' distribution is relatively straightforward.

Another dominant approach to quantitative text analysis is (manual or automatic) **annotation**. Annotating a text has many commonalities with tagging in biology. Instead of manipulating the micro-structures (e.g., genome) of a text (whatever that is), linguists annotate each expression with meta-expressions representing some features. For instance, (4a) can be annotated with grammatical features (e.g., part of speech) as in (7)[10]. Assigned word classes (e.g., VERB, NOUN) enable an analyst to conduct a more detailed analysis. If an analyst decides to investigate the kind of words occurring around verbs, naive collocation analysis does not suffice because naive collocation analysis only deals with actual words (e.g., *run*, *runs*, *running*).

(7) I PROPER NOUN ran VERB back ADVERB and CONJUNCTION turned VERB him PRONOUN, ever ADVERB so ADVERB gently ADVERB, to INFINITIVE get VERB at PREPOSITION his DETERMINER hands NOUN.

[9] For simplicity, we only discuss co-occurring adjectives.

[10] Since (7) employs a naive word class system for simplicity, one should not treat (7) as a valid analysis.

The keystone of feature analysis is a consistent annotation strategy. In the part-of-speech tagging (i.e., word class annotation) like (7), they usually employ a more sophisticated, finer-grained strategy. Since word classes are defined in terms of their formal distributional patterns [17, pp.29–32], they are relatively stable. For instance, as shown in (8), possible words succeeding the proper noun "Nana" are classified as VERB, and others are not[11]. One can employ a more thorough context to delineate each word class.

(8) Nana _____.

a. i. Nana barked.
   ii. Nana ran.

b. i. * Nana fortunately.
   ii. * Nana luckily.

In principle, analysts have the freedom to assign any kind of annotations. For simplicity, we only dealt with grammatical features, but it is totally possible to annotate each sentence with semantic (or even pragmatic) information. In analysing *run*, Gries annotated various senses to see how each sense is differentiated depending on various types of variables [20]. Following Gries, Glynn attempted to replicate Gries' analysis by employing the same strategy and succeeded it [29].

Text annotations advance the (practical and theoretical) science of language. Though it is challenging to devise a consistent annotation strategy for semantic information, some attempts have been made [1, 2, 9]. These valuable resources are essential in acquiring semantic information from text. In corpus linguistics, it is almost impossible to analyse anything without dealing with contents expressed by meta-language(s).

3. Operationalising meaning: Specificity and granularity

Before tackling the pragmatic constraints in research processes of quantitative linguistics, this section introduces two criteria in concept structuring: specificity and granularity. Roughly put, the former corresponds to class inclusion (i.e., DOG $\subset$ ANIMAL), and the latter to part-whole relation (i.e., FINGER $\sqsubset$ HAND). Though many scholars, including philosophers and linguists, have attempted to provide a framework to describe concepts [18, 38], these two criteria are not avoidable in any descriptions. As will be clear in Section 4, these two dimensions of meaning contribute to the understanding of research practices in quantitative linguistics.

Since the dawn of linguistic semantics, semantic relations between words have been treated as one of the crucial aspects of meaning [7]. If we take two random

[11] In (8), an asterisk "*" signals an unacceptable sentence.

words a number of times, we are likely to find many semantic relations. Some of the most well-known relations in linguistics are (i) hyponymy (i.e., *dog : animal*)[12], (ii) meronymy (i.e., *hand : arm*), (iii) synonymy (i.e., *sofa : couch*), and (iv) antonymy (i.e., *tall : short*). Fellbaum distinguishes these relations into two categories: lexical and conceptual-semantic relations [10, pp.351–352]. The former corresponds to the relation between words, and the latter to the relation between concepts. Like many lexical semanticists [48, p.123], Fellbaum categorised (iii) synonymy and (iv) antonymy as lexical relations and (i) hyponymy and (ii) meronymy as semantic-conceptual.

Semantic-conceptual relations are indispensable in describing discrete concepts. For instance, one may characterise the concept, Dog as a subcategory of Animal, while others may characterise it as a collection of Paw, Leg, Trunk, Muzzle, etc. These two criteria are independent but complementary[13]. Though hyponymy can be defined in a naive set-theoretic way, meronymy is not.

Let Dog be a class (or set) of dogs and Animal be a set of animals. It follows that all members of Dog also belong to Animal, meaning that Dog is a subclass (or subset) of Animal. Meronymy is more challenging [60]. Let Finger be a set of (human) fingers and Hand be a set of (human) hands. Though it seems evident that Finger is part of Hand, each set contains different types of things, meaning that the relation between Finger and Hand cannot be characterised in a naive set-theoretic fashion. We will not discuss the mathematical characterisations of meronymy in this paper. Still, we use $x \subset y$ to signal the class inclusion relation between x and y (e.g., Dog $\subset$ Animal) and $x \sqsubset y$ to signal part-whole relation between x and y (e.g., Finger $\sqsubset$ Hand) for notational convenience.

An analyst must decide the accuracy levels in describing (or identifying) concepts. For instance, one could describe a fluffy-looking friendly creature as Golden Retriever, or Animate Entity. We call this degree of concept identification as **specificity**, which corresponds to the choice of words in the hierarchy of hyponymy. Once an analyst decides the specificity of a concept, she can determine the level of **granularity**. There is a trade-off relation between specificity and granularity: The more specific the concept is, the easier it gets to delineate its parts. For instance, parts of Golden Retriever are more accessible than those of Animate Entity.

These two criteria are treated as crucial distinctions in **ontological engineer-**

[12] Semantically (syntactically, or, morphologically) related lexical items are notated as $x : y$ for convenience. We give specific notations to hyponymy and meronymy in the following discussion.

[13] In frame semantics [13, 15], lexical items are described relative to **(semantic) frames** which are defined in terms of part-whole relationship and class inheritance. For the reason of space, we do not discuss details of frame semantics, but two criteria explained in Section 3 are exploited fully in frame semantics.

ing, an interdisciplinary field which aims to construct a rigid characterisation of concepts for artificial intelligence [43, 44, 45]. In computer science, hyponymic relations are referred to as IS-A relations and meronymic relations as PART-OF relations[14]. Following the discussion by Mizoguchi [45, pp.194–203], we assume that hyponymic relations are inseparable from members' identities, while meronimic relations are not. If an entity named "Nana" is a DOG, she must be born as DOG, and die as a member of DOG. By contrast, if an entity is a FINGER, it does not always have to be realised as a part of HAND.

In linguistic semantics (even in linguistic pragmatics), a linguist aims to describe and explain the meaning of target expressions. No matter what kind of descriptive frameworks she employs, she has to decide the right degree of specificity and granularity of the meaning (or concept). For instance, senses in (4) were labelled differently and characterised in certain degrees of specificity and granularity. Each sense had its parts in a certain degree of specificity. In principle, one could argue that all senses in (4) correspond to RUNRELATEDSENSE, but it would be pointless to describe a few commonalities observed in these sentences.

4. Pragmatic constraints in corpus linguistic research

As the discussion in 3 suggests, corpus linguists must decide the degrees of specificity and granularity in identifying or describing a concept. In every empirical enquiry, a researcher must determine the kind of phenomenon to investigate and explicitly state how the target is observed. Without such procedures, replicating one's analysis becomes impossible. The degrees of specificity and granularity are vital in effective communication between analysts. We refer to specificity and granularity as **precision** for convenience.

For instance, let's say that an analyst A could argue that there is only one kind of sentence in English by setting the lowest degree of precision. At the same time, another analyst B could discuss infinite types of sentences in English by setting the highest degree of precision. Both A and B are right and wrong. A would have difficulty identifying the patterns shared by all members of English sentences. By contrast, B can easily generalise the target expressions since relatively few sentences are analysed.

In principle, approaches taken by A and B are possible but not probable because extreme degrees of precision lead to miscommunication. If we were to take every sentence belonging to only one class, we would not be able to account for the

[14] Technically, IS-A relations and PART-OF relations only hold among concepts, while hyponymy and meronymy hold among (typically) words. Though we ignore this distinction for simplicity, it is critical to distinguish the kinds of entity that can stand in such relations.

heterogeneity among the members. Likewise, if we were to take every sentence belonging to a different class, we would not be able to account for the commonalities across different classes. Modulation of precision is vital in selecting target expressions because it is an inevitable step in conducting empirical research.

Modulation of precision is also pervasive in corpus linguistic analyses, which is affected by **pragmatic constraints**. Pragmatic constraints refer to the various constraints that interfere with scientific processes to increase communicativity among researchers, or understandability by other researchers. Since the right degree of precision cannot be decided in an a priori fashion, analysts must explore the agreeable scale, which must be carried out pragmatically.

This section is structured as follows. Section 4.1 and 4.2 discuss how pragmatic modulation of precision comes into play in feature and collocation analysis, respectively. We limited our discussion involving the problem of meaning for simplicity.

4.1. Pragmatic constraints in collocation analysis

As introduced in Section 2.2.1, analysts attempt to find frequent patterns of words in collocation analysis. As Table 1 shows, collocation analysis can be conducted objectively, which means the result is easily replicated by sharing a series of procedures. However, one of the major challenges in collocation analysis is to set a range of observations. We illustrate this point by exploiting specificity and granularity, defined in Section 3.

To investigate collocations of the target expressions, linguists decide the specificity of target expressions. Firstly, they decide how specific (or concrete) the target expression is. For instance, if a linguist chooses to investigate the collocations of the word *dog*, she has the choice of whether she includes the singular form "dog" along with the plural form "dogs". Morphologically related forms of a word are called **word forms**, and their root expressions **lemma**. Relations between lemmas and their word forms are a matter of specificity. Linguistically, "dogs" and "dog" are kind of *dog* because the semantics of each word form coincide much with *dog*.

In addition to the matter of specificity, the choice of granularity is involved in collocation analysis. When linguists examine the collocational patterns of their choice, they rarely investigate the word next to the target expressions. For instance, the left words ending with the word "dog" in BNC using Sketch Engine [35][15] are "the dog" (2,239), "a dog" (1,258), "your dog" (272), "his dog" (186), and "of dog" (136). This result is not as "interesting" as one may want it to be because the word "dog" is a countable noun, and it has to occur with determiners (e.g., *a*, *the*, *his*, *your*) due to one of the grammatical constraints in English speaking community. The range of observation corresponds to the matter of granularity. In identifying the character-

[15] The frequency list of words is omitted in this paper. However, it is available on URL

istics of linguistic phenomena, linguists must decide the "right" degree of range of co-occurring words.

This section is structured as follows: Section 4.1.1 explores pragmatic constraints in collocation analysis, and Section 4.1.2 in feature analysis.

4.1.1. Modulation of specificity in collocation analysis

This section discusses the pragmatic constraints in selecting the degree of specificity. We argue that accepting unjustified conventional assumptions in corpus linguistic research leads to avoiding conflicts among linguists by reviewing the part of the result by Gries [23]. The selectional process affected by the convention follows the pragmatic constraints described above.

Gries explored how the choice of specificity affects the analysis through the quantitative analysis of the ditransitive construction (i.e., double object construction), as exemplified in (9)[16]. The ditransitive construction is often represented as a construction with one subject (which instantiates the "actor" of object transfer (i.e., AGENT)) and two objects (which instantiate the "object" of transfer (i.e., THEME) and the "receiver" (i.e., RECIPIENT)).

(9) [SUBJECT$_{\text{AGENT}}$ VERB OBJECT$_{\text{RECIPIENT}}$ OBJECT $_{\text{THEME}}$]

a. He gave her the book.

b. She told him a story.

[23, p.241]

Degrees of specificity depends heavily on analysts' (often conventional) choices. Determining the "right" degree of precision in the linguistic analysis is rarely discussed. However, Gries [23] is one of the few exceptions. Gries contrasted (i) lemma- and word form- based analyses and (ii) register-based analysis to observe how results vary in each condition. The former corresponds to the selection of specificity, and the latter to the choice of granularity. Below, we review the consequences of such selections by examining Gries' results. Gries explored them through the quantitative analysis of the ditransitive construction (i.e., double object construction), as exemplified in (9)[17]. The ditransitive construction is often represented as a construction with one subject (which instantiates the "actor" of object transfer (i.e., AGENT)) and two objects (which instantiate the "object" of transfer (i.e., THEME) and the "receiver" (i.e., RECIPIENT)).

[16] Note that the schematic representation of the construction in (9) are not the same as the original. This change is strictly for simplicity and does not affect the Gries' discussion.

[17] Note that the schematic representation of the construction in (9) are not the same as the original. This change is strictly for simplicity and does not affect the Gries' discussion.

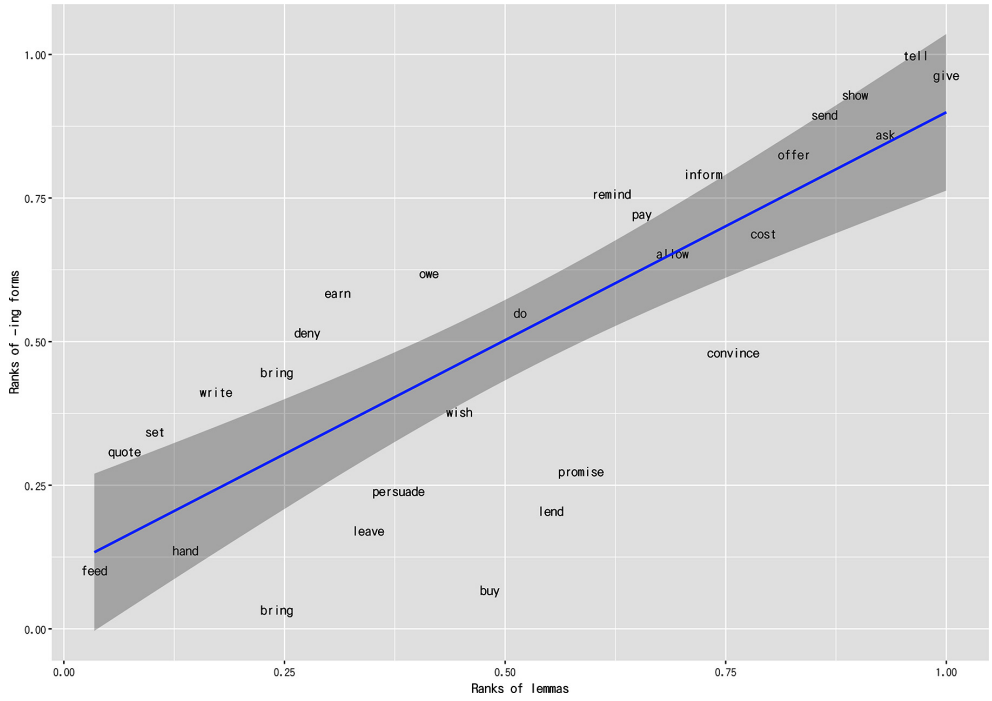

Figure 2 Correlation of ranks between lemmas and gerunds (A simplified, recreated version of [23, p.245])

In principle, the characterisation of ditransitive constructions can be carried out on any degree of specificity. In the linguistics circle, lemma-based analyses are dominant, which is often unjustified. As Gries [23, p.238–241] points out, some studies challenge this widely held assumption. Gries extracted all ditransitive constructions from ICE-GB corpus[18] and annotated all word forms. An English verb has four inflectional forms: present, past, present participle, and past participle. Gries compared the ranks of verbs in all four groups with that of the lemma and concluded that inflectional forms might affect the qualitative findings. In contrast, they do not affect the overall quantitative tendency. Figure 2 visualises the relations of ranks between lemma and present participle forms[19]. The x-axis represents the lemmas' ranks (converted into ratio variable), and the y-axis the present participle (-ing) forms' ranks.

[18] `https://www.ucl.ac.uk/english-usage/projects/ice-gb/`

[19] This plot was created by the first author using R [55] on the Gries' data. Due to the limitation of data, only the relation between lemmas and -ing forms was analysed. Moreover, Gries [23] employs the results of **collostructional strength** [24]. Though it is preferable to repeat the process, the limitation of the original data prevented from replicating. For this reason, Figure 2 only deals with percentiles of raw frequencies.

As the scatter plot in Figure 2 shows, the overall quantitative tendency in lemma- and word form-based analysis is relatively stable. Linguists tend to choose a lemma-based study for this construction over the other. The choice of specificity in collocation analysis is usually affected by the conventions since it is not practical to assume that inflectional forms cause (statistically or linguistically) significant differences. Such attitude is motivated by the textbook treatment of inflectional forms. For instance, Fromkin [17, p.34–35] introduces **derivational** morphology (opposed to inflectional morphology) as a means to increase vocabulary in a language and allow speakers to express things differently (e.g., *establish : establishment*). This is to say, unlike derivational morphology, inflectional morphology contributes (relatively) little to meanings.

Often, corpus linguistics is portrayed as an empirical enterprise rather than a rational one. As discussed in Section 2, corpus linguistics was born from resisting rationalistic approaches to meaning. As Fillmore emphasises [11], corpus linguistics provides an objective way to observe linguistic usages. For this reason, linguists devoted to corpus methods tend to consider themselves empiricists [21]. However, as the choice of specificity suggests, such linguists do not always question every aspect of generalisations from theoretical investigations. Though it is difficult to lay out the kind of assumptions worth investigating in corpus linguistics, some assumptions are held without serious empirical grounds. Corpus linguists can doubt every assumption in linguistics if they wish to. However, this is not a practical goal to pursue. As the selection of specificity suggests, corpus linguists also accept some assumptions without empirical grounding. In analysing the characteristics of ditransitive constructions, how corpus linguists choose specificity is mainly affected by the conventions in the linguistics circles.

As Quine discusses [53, 54], scientists do not discard every assumption whenever they encounter a phenomenon that contradicts traditional assumptions. Instead of discarding the whole theory, they gradually change their assumptions so that researchers can communicate with each other. As suggested, corpus linguists can question all assumptions in (mainly theoretical) linguistics. However, it can lead to miscommunications with other linguists. For this reason, corpus linguists tend to accept some basic assumptions held conventionally. In this sense, the choice of specificity in ditransitive constructions is interesting because the selectional process in the analysis is questioned, which is caused by the emergence of usage-based conceptions in linguistics.

Theories and their practices in linguistic enquiry work in tandem. Though it is ideal for giving empirical grounding to every theoretical assumption, such attitudes can lead to miscommunication among linguists. To avoid such conflicts, corpus linguists pick the "right" degree of specificity to discuss the characteristics of the target expressions. Gries' analysis was fruitful due to its rising interest in the currents of

theoretical linguistics. Factors involved in a research program are constrained pragmatically.

4.1.2. Modulation of granularity in collocation analysis

This section discusses the pragmatic constraints in the modulation of granularity. In collocation analysis, linguists observe frequent collocations around the target expressions. However, the range of observations is also open to discussion. We see that corpus linguists tend to determine the appropriate range, which allows them to understand the characteristics of the target.

The discussion in Section 4.1.1 identified ditransitive constructions by their verbs (i.e., *tell, give, show, ask, send*). The other important factor in collocation analysis is the range of observation employed in research. For instance, adding extra adjuncts to most English sentences is relatively easy, as demonstrated in (10). The data in (10) suggests that adjuncts are nothing but optional elements since deleting additional elements in (10b–d) does not yield an unacceptable sentence as exemplified in (10a).

(10) a. Alice gave Bill the book
b. Alice gave Bill the book in Japan.
c. Alice gave Bill the book in Japan for his birthday.
d. Alice gave Bill the book in Japan for his birthday in January.

In principle, linguists can choose any range of collocations. Many linguists conduct quantitative analysis with the help of corpus interfaces. In such systems, linguists can select the range of collocations to observe. For instance, Sketch Engine [35], one of the most effective corpus interfaces available, offers a means to obtain the frequency list of two to six collocations around the target expressions. However, like in the case of specificity of constructions, linguists tend to select such ranges without seriously contemplating their groundings.

Suppose that a corpus linguist decides to analyse the difference between the ditransitive construction with *give* and *tell*. If she analyses them in terms of their collocations, it is necessarily straightforward to determine the appropriate range for the investigation. Often she investigates two to five words around the target expressions (i.e., *give*, *tell*). However, it is impossible to know the appropriate range to reveal the difference of the target expressions in an a priori fashion. Linguists have no choice but to determine such a range in a bottom-up manner.

Theoretically, it is possible to assume that any range of collocations can contribute to the characterisation of the target expressions. In natural language processing (NLP, for short), semantic knowledge is obtained automatically using collocational information [5]. For instance, the semantics of a word is equated with collocation. Figure 3 is a schematic representation of *orange* and *fruit*. In Figure 3, each bar

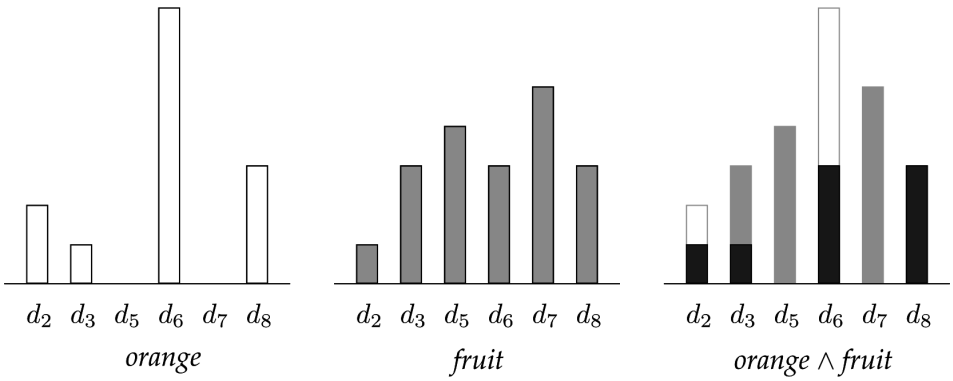

Figure 3 Vector representations of the terms [5, p.49]

corresponds to the occurrence of each word in a document (i.e., $d_2, d_3, \ldots$). These schematic "documents" can be equated with collocations of each word (e.g., "Alice likes _____"). By equating collocational tendency with the semantics of given words, she can compute each word's (dis)similarity.

This strategy is computationally sound and has proven helpful in many applications. However, the issue lies in the kind of collocations linguists deal with. Figure 3 shows six vectors to represent each word, but how we determine the "significant" collocation is not yet clear. An automated word vector machine can obtain such information by examining all words. In contrast, humans cannot conduct such investigations. Especially in corpus linguistic research, one has to determine the "moderate" range of observation. For instance, a machine programmed to obtain unique collocational patterns for every word in a corpus does not "care" if the significant collocation is found as forty-six words long (the number is arbitrary).

Again, such decisions are constrained pragmatically. In collocational analysis, linguists tend to look for significant patterns in a moderate range (usually less than ten words), which could possibly yield an insignificant result. However, it is tricky for linguists to discuss the extremely long collocations as indicators of the target's meaning. Corpus linguists do not aim to discover the significant collocations but to reveal the "interesting" quantitative tendency of natural language. Interpreting a series of frequent collocations requires careful inspection, and the result must be easy enough for human scholars to understand.

4.2. Pragmatic constraints in feature analysis

This section discusses the pragmatic constraints in feature analysis[20]. Assigning a metalanguage that captures some characteristics to an object language is common

[20] Portions of this section were cited from Kambara et al. [34].

in corpus linguistics. Though annotation with formal linguistic features (e.g., part of speech) is relatively easy, that with semantic features is quite challenging due to the "ungraspability" of meanings, which has been recognised in academic circles for some decades. Despite the widely-acknowledged difficulty, building a framework for semantic annotation has been attempted [9, 13].

Two kinds of semantic annotation are currently available: (i) semantic class labelling and (ii) semantic role labelling. The former assigns the specific semantic class to a given text (e.g., assignment of DOG to the expression *"Nana"*) while the latter assigns the role in a situation to a given text (e.g., assignment of ⟨AGENT⟩ to the expression *"Nana"* in *"Nana hid a bone in the garden"*). These two criteria are independent but complementary of each other [2]. For the reasons explained in detail in 4.2.2, the former corresponds to the matter of specificity and the latter to the matter of granularity. Section 4.2.1 discusses the semantic class annotation and 4.2.2 the semantic role annotation.

4.2.1. Modulation of specificity in feature analysis

This section discusses the pragmatic constraints on modulating the degree of specificity in semantic class labelling. The specificity of annotation corresponds to generality of semantic information. As well known in the set theory, the richness of intension correlates with the poverty of extension, and the richness of extension correlates with the poverty of intension. In semantic annotation, linguists must consider the ontological nature of each category. Otherwise, they confront with tough generalisations.

In assigning a semantic class to a given text, linguists must decide the kind of ontology they employ. For instance, examples in (11) differ in each grammatical subject. Their acceptability is clearly affected by the semantic class of each subject. (11a) is natural because the subject is a kind of DOG and (11b) is not because the subject is a kind that does not "bark"[21]. To capture such characteristics, linguists may annotate (11) as (12) in which subjects are annotated with respectable semantic classes (i.e., DOG, HUMAN).

(11) a. The golden retriever barked.
b. ? The teacher barked.

(12) a. The golden retriever $_{\text{DOG}}$ barked.
b. ? The teacher $_{\text{HUMAN}}$ barked.

Although semantic class labelling seems quite straightforward, linguists have the

[21] Like many verbs in natural language, *bark* is polysemous. (11b) is natural if one interprets "barked" as *"to utter in shouting tone"*. For simplicity, we assume that "barked" refers to the action of crying by dogs.

freedom to choose any degrees of specificity. If they wish, they can implement a full-fledged-biologically-sound system for semantic class annotation. However, such attempts would be improbable because applicability of scientific (or biological) classification in linguistic semantics is not justified, and linguistic semantics are affected by folk biology.

In anthropological linguistics, (linguistic) semantic categories are well studied [4][22]. According to Murphy [48, p.115–116], the specificity of categories are motivated by the kind of speakers' lives. Since members of scientific communities aim to discover the kind of things in the world, they have implemented the scientific classification system. On the other hand, groups of non-scientists do not necessarily share the same goal. For instance, it is not surprising if a speaker of some linguistic community categorises a bat as a kind of bird. Since natural language is used by many agents with different goals, its foundation cannot be based on a particular community.

Re-constructing a classification system for a linguistic semantics requires a labour intensive work. For this reason, when a corpus linguist annotates a given text, they tend to recycle the folk-biological classification system. However, the "right" degree of specificity cannot be determined independent of the target expressions. For instance, to capture the difference of acceptability of examples in (11), HUMAN and DOG were required. In contrast, underlined expressions in (13) causes (un)acceptability in relation to their animacy (i.e., what is alive).

(13) a. Alice let her students HUMAN sleep
b. Alice let her dogs DOG sleep
c. ? Alice let her books ARTEFACT sleep

Classification systems must be coherent, and effective in describing the different characteristics of many different phenomena. Especially in a quantitative corpus linguistic research, how such systems are employed is not necessarily clear because of the multifactorial nature of language. Even when linguists only focus on semantic classification, they need to devise the "right" degree of specificity in their enquiries. Such processes are often supplemented by consulting sources like thesauri (e.g., WordNet [9]). With or without such resources, corpus linguists tend to identify such a degree by observing actual sentences and convince their colleagues of the appropriateness in modulating the specificity, which can only be conducted pragmatically. As long as the proposed degree of specificity captures the characteristics of their target expressions, they would not have problems. However, in a large scale annotation project, such characteristics would be too general to capture by a coarse-grained semantic classification system.

[22] See Murphy [47, p.69–74] for a brief overview of the anthropological approach.

4.2.2. Modulation of granularity in feature analysis

This section discusses the modulation of granularity in semantic role labelling. Unlike semantic class labelling, semantic role labelling is a description of the event's participants. Depending on the specificity of events, the granularity of events varies. In assigning appropriate semantic roles to a given text, corpus linguists must carefully determine the kind of events so that the results do not lead to miscommunication.

One of the goals in linguistic semantics is to identify and explain distributions of **semantic roles** in a given sentence [37, Ch.5]. Descriptions of semantic roles specify "Who did What to Whom, and How, When and Where?" in a given sentence [49, p.2]. The sentences in (14) are annotated with some of the typical semantic roles: ⟨AGENT⟩, ⟨THEME⟩, and ⟨INSTRUMENT⟩. Roughly, ⟨AGENT⟩ refers to "the actor" of the situation expressed by the predicate, ⟨THEME⟩ to "the influenced entity", and ⟨INSTRUMENT⟩ to "the entity used by the actor". Assigning semantic roles to syntactic constituents is not necessarily straightforward since the manners of their distribution vary from context to context. As exemplified in (14), the different semantic roles are assigned to the syntactically same constituents of the sentences with the verb *open*.

(14) a. John ⟨AGENT⟩ opened the door ⟨THEME⟩.
b. The door ⟨THEME⟩ was opened by John ⟨AGENT⟩.
c. The key ⟨INSTRUMENT⟩ opened the door ⟨THEME⟩.
d. John ⟨AGENT⟩ opened the door ⟨THEME⟩ with the key ⟨INSTRUMENT⟩.

[12, p.59]

Although determining the appropriate semantic roles is tough, some widely-recognised semantic roles can be summarised as Table 2[23] which contains some roles that have already been explained above. In principle, the semantic roles in the table allow linguists to annotate the kind of events expressed by the given text. Depending on verbs (or sentences as observed in (14)), the semantic interpretations of grammatical constituents can vary. For instance, unlike the verb *run*, the subject of the verb *sleep* must be interpreted as ⟨EXPERIENCE⟩ because the verb *sleep* denotes one's state of sleeping, which does not involve an active engagement in an event.

Since semantic roles denote the characteristics of an event's participant, identification of semantic roles can be equated with that of events. For this reason, semantic roles in Table 2 are realised in groups (e.g., {⟨AGENT⟩, ⟨THEME⟩, ⟨INSTRUMENT⟩}, {⟨INSTRUMENT⟩, ⟨THEME⟩}), as shown in (14). Especially in a large scale semantic

[23] Some notations of Table 2 is modified. The change does not affect the contents of the table.

Table 2 A set of widely recognised semantic roles [49, p.4]

Role	Description	Examples
⟨AGENT⟩	Initiator of action, capable of volition	**The batter** smashed the pitch into left field. **The pilot** landed the plane as lightly as a feather.
⟨PATIENT⟩	Affected by action, undergoes change of state	David trimmed **his beard**. John broke **the window**.
⟨THEME⟩	Entity moving, or being "located"	Paola threw **the Frisbee**. **The picture** hangs above the fireplace.
⟨EXPERIENCER⟩	Perceives action but not in control	**He** tasted the delicate flavor of the baby lettuce. **Chris** noticed the cat slip through the partially open door.
⟨BENEFICIARY⟩	For whose benefit action is performed	He sliced **me** a large chunk of prime rib, and I could hardly wait to sit down to start in on it. The Smiths rented an apartment **for their son**.
⟨INSTRUMENT⟩	Intermediary/means used to perform an action	He shot the wounded buffalo with **a rifle**. The surgeon performed the incision with **a scalpel**.
⟨LOCATION⟩	Place of object or action	There are some real monsters hiding in **the anxiety closet**. The band played on **the stage**.
⟨SOURCE⟩	Starting point	The jet took off from **Nairobi**. We heard the rumor from **a friend**.
⟨GOAL⟩	Ending point	The ball rolled to the other end of **the hall**. Laura lectured to **the class**.

annotation project, assigning appropriate semantic roles can be challenging because the specificity of events must be determined before assigning semantic roles. Deciding the appropriate specificity of events can be tricky because linguists must agree to employ the specific event(s).

Many problems arise in assigning every sentence with the limited number of roles [37, Ch.5][24]. For instance, semantics of some verbs is not affected by the order of subjects and objects. As exemplified in (15), the verb *see* clearly specifies "Who did What" as its subjects and objects, while the verb *resemble* does not. Though

[24] Another problem with assuming the limited number of roles is related to the specificity of roles. For instance, ⟨BENEFICIARY⟩ can be treated as a kind of ⟨PATIENT⟩. To recognise this inclusion relationship, linguists must devise an ontology of events. Kambara et al. [34] discusses the strength of frame semantics in this regard.

the possible combinations of semantic roles are quite large in number, they are not sufficient in describing a variety of contents expressed by natural language.

(15) a. i. Alice saw Charlotte
ii. Charlotte saw Alice
b. i. Alice resembles Charlotte
ii. Charlotte resembles Alice

Since recognising the appropriate parts of an event is extremely challenging, (corpus) linguists tend to resort to employing the representative roles in Table 2. The choice of semantic roles imposes the kind of event ontology linguists share. However, linguists make use of currently available roles since it is not practical to implement a full-fledged event ontology. The process of developing a new system for semantic role annotation is highly pragmatical. For instance, Fillmore et al. [14] discusses how to implement a set of semantic roles for the verb *attach* in detail. In the process, they describe the target in relation to conventionally available resources (e.g., dictionaries). A gradual process of characterising the semantic roles concords with the analogy of Neurath's boat by Quine [53].

The pragmatic nature of semantic role labelling is apparent due to the "ungraspability" of meanings. Though many scholars tackled the meanings, the conception of the grand theory of meaning is highly controversial to say the least. Different theories have different goals, which makes corpus linguists challenging to determine the "right" theory in tasks they confront. For linguists to employ a brand new theory, it must be appealing for the kind of meaning they analyse. Pragmatically speaking, it is improbable to utilise a completely unfamiliar theory in their familiar tasks. For this reason, they resort to employing the "good-old" lists of semantic roles.

5. Conclusion

This paper aimed to demonstrate the ubiquitous nature of pragmatic constraints in corpus linguistics. We sketched a system to describe a concept by introducing the degrees of precision. In applying the framework in the process of corpus linguistic enquiries, we constantly confirmed a tendency to appeal to conventions or communicativity of such concepts. As long as corpus linguistic methodologies are driven by human beings like other scientific disciplines, pragmatic constraints are inevitable. When a linguist decides to analyse linguistic semantics, she does not have access to the one and only true meaning. Instead, she must resort to whatever resources she has, which leads to the rejection of Platonistic meaning, which many pragmatists have argued.

Several issues remain unsolved. First, a more detailed characterisation of pragmatic constraints is needed. Crucial aspects of these constraints revolved around

the "interestingness" and "communicativity between researchers". These factors still need to be clarified. Secondly, we employed two criteria in concept structuring: specificity and granularity. The degrees of precision in the philosophy of science has yet been explicated enough.

References

[1] C. F. Baker, FrameNet: Frame semantic annotation in practice. In N. Ide and J. Pustejovsky (eds.), Handbook of Linguistic Annotation (pp. 771–811). New York: Springer, 2017.
[2] C. F. Baker and C. Fellbaum, WordNet and FrameNet as complementary resources for annotation. In Third Linguistic Annotation Workshop (pp. 125–129), 2009.
[3] R. J. Bernstein, The Pragmatic Turn. Cambridge: Polity, 2010.
[4] B. Berlin, D. E. Breedlove, and P. E. Raven, General principles of classification and nomenclature in folk biology, American Anthropologist 75, 214–242, 1973.
[5] D. Clarke, A context-theoretic framework for compositionality in distributional semantics, Computational Linguistics 38(1), 4171, 2012.
[6] N. Chomsky, Aspects of the Theory of Syntax. Cambridge, Mass.: MIT Press, 1965.
[7] A. D. Cruse, Lexical Semantics. Cambridge: Cambridge University Press, 1986.
[8] A. D. Cruse, Meaning in Language: An Introduction to Semantics and Pragmatics. Oxford: Oxford University Press, 2011.
[9] C. Fellbaum (ed.), WordNet: An Electronic Lexical Database. MIT Press, 1998.
[10] C. Fellbaum, Lexical relations. In J. R. Taylor (ed.), The Oxford Handbook of the Word (pp. 350–363). Oxford: Oxford University Press, 2015.
[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 (pp. 35–66), Berlin: Mouton de Gruyter, 1992.
[12] C. J. Fillmore, Form and Meaning in Language, Vol.1: Papers on Semantic Roles. Stanford: CSLI, 2003.
[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, M. Petruck, J. Ruppenhofer, and A. Wright, FrameNet In action: The case of Attaching. International Journal of Lexicography 13(3), 297–332, 2003.
[15] 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 (pp. 791–816). Oxford: Oxford University Press.
[16] J. R. Firth, A synopsis of linguistic theory. In F. R. Palmer (Ed.) Selected Papers of J. R. Firth 1952–59 (pp. 1–31). London: Longmans, 1968.
[17] V. A. Fromkin (ed.), Linguistics. London: Blackwell, 2000.
[18] T. Gamerschlag, D. Gerland, R. Osswald, and W. Petersen (eds.), Frames and Concept Types: Applications in Language and Philosophy. New York: Springer, 2013.
[19] D. Geeraerts, The doctor and the semantician. In D. Glynn, and K. Fischer (eds) Quantitative Methods in Cognitive Semantics: Corpus-driven Approaches (pp. 63–78). Berlin: Mouton de Gruyter, 2010.

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

[21] S. Th. Gries, Corpus linguistics and theoretical linguistics: A love-hate relationship? Not necessarily... International Journal of Corpus Linguistics, 15(3), 327–343, 2010.

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

[23] S. Th. Gries, Corpus data in usage-based linguistics: What's the right degree of granularity for the analysis of argument structure constructions? In M. Brdar, S. Th. Gries, and M. Ž. Fuchs (eds.), Cognitive Linguistics: Convergence and Expansion (pp. 237–256). Amsterdam: John Benjamins, 2011.

[24] S. Th. Gries, More (old and new) misunderstandings of collostructional analysis: on Schmid & Küchenhoff (2013). Cognitive Linguistics, 26(3), 505–536, 2015.

[25] S. Th. Gries, Quantitative Corpus Linguistics with R: A Practical Introduction. Routledge, 2016.

[26] S. Th. Gries, Quantitative corpus methods of cognitive semantics/linguistics. In Thomas Li (ed.), Handbook of Cognitive Semantics (pp.328–350). Cologne: Brill, 2022. `https://www.stgries.info/research/ToApp_STG_QuantCorpCognSem_HdbkCognSem.pdf`

[27] S. Th. Gries, and D. S. Divjak, Quantitative approaches in usage-based cognitive semantics: Myths, erroneous assumptions, and a proposal. In D. Glynn, and K. Fischer (eds.), Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches (pp. 333–353). Berlin: Mouton de Gruyter, 2010.

[28] D. Glynn, Techniques and tools: Corpus methods and statistics for semantics. In D. Glynn and J. A. Robinson (eds.), Corpus Methods for Semantics (pp. 307–341). Amsterdam: John Benjamins, 2014.

[29] 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 (pp. 117–144). Amsterdam: John Benjamins, 2014.

[30] P. Godfrey-Smith, Quine and pragmatism. In G. Harman and E. Lepore (eds.), A Companion to W.V.O. Quine (pp. 54–68). New Jersey: Wiley, 2014.

[31] A. E. Goldberg, Constructions: A Construction Grammar Approach to Argument Structure. Chicago: University of Chicago Press, 1995.

[32] M. Jakubíček, A. Kilgarriff, V. Kovář, P. Rychlý, and V. Suchomel, The TenTen corpus family. In 7th International Corpus Linguistics Conference CL 2013 (pp. 125–127), 2013.

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

[34] K. Kambara, H. Nozawa, and T. Takahashi, Toward a finer-grained specification of frames: A corpus-based approach to argument realisation, submitted.

[35] 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, 7–36, 2014.

[36] K. Kuroda, J. Kazama, and K. Torisawa, A look inside the distributionally similar terms. In Proceedings of the Second Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010) (pp. 40–49), 2010.
[37] B. Levin, and M. Rappaport Hovav, Argument Realization. Cambridge: Cambridge University Press, 2005.
[38] S. Löbner, T. Gamerschlag, T. Kalenscher, M. Schrenk, and H. Zeevat (eds.), Concepts, Frames and Cascades in Semantics, Cognition and Ontology. New York: Springer, 2021.
[39] J. D. McCawley, Everything that Linguists have Always Wanted to Know about Logic ... But Were Ashamed to Ask. Chicago: University of Chicago Press, 1993.
[40] J. D. McCawley, The Syntactic Phenomena of English. Chicago: University of Chicago Press, 1998.
[41] T. McEnery, and H. Andrew, Corpus Linguistics: Method, Theory and Practice. Cambridge: Cambridge University Press, 2012.
[42] T. McEnery, and V. Brezina, Fundamental Principles of Corpus Linguistics. Cambridge: Cambridge University Press, 2022.
[43] R. Mizoguchi, Tutorial on ontological engineering Part 1: Introduction to ontological engineering. New Generation Computing, 21(4), 365–384, 2003.
[44] R. Mizoguchi, Tutorial on ontological engineering Part 2: Ontology development, tools and languages. New Generation Computing, 22(1), 61–96, 2004.
[45] R. Mizoguchi, Tutorial on ontological engineering Part 3: Advanced course of ontological engineering. New Generation Computing, 22(2), 193–220, 2004.
[46] J. P. Murphy. Pragmatism: From Peirce to Davidson. Colorado: Westview Press, 1990.
[47] M. L. Murphy, Semantic Relations and the Lexicon: Antonymy, Synonymy, and Other Paradigm. Cambridge: Cambridge University Press, 2003.
[48] M. L. Murphy, Lexical Meaning. Cambridge: Cambridge University Press, 2010.
[49] M. Palmer, D. Gildea, and N. Xue, Semantic Role Labeling. California: Morgan & Claypool Publishers, 2010.
[50] C. S. Peirce, Some consequences of four incapacities. Journal of Speculative Philosophy, 2(3), 140–157, 1868.
[51] C. S. Peirce, The fixation of belief. Popular Science Monthly, 12, 1–15, 1877.
[52] C. S. Peirce, How to make our ideas clear. Popular Science Monthly, 12, 286–302, 1878.
[53] W. V. Quine, Word and Object, Cambridge, Mass.: MIT Press, 1960.
[54] W. V. Quine, From a Logical Point of View: 9 Logico-Philosophical Essays. New York: Harper & Row Publishers, 1961.
[55] R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2022. URL: `https://www.R-project.org/`.
[56] P. Redding, Feeling, thought and orientation: William James and the idealist anti-Cartesian tradition. Parrhesia 13, 41–51, 2011.
[57] R. Rorty, Objectivity, Relativism, and Truth: Philosophical Papers, Volume 1. Cambridge: Cambridge University Press, 1990.
[58] A. Stefanowitsch, Corpus Linguistics: A Guide to the Methodology. Berlin: Language

Science Press, 2020.
[59] L. Talmy, Toward a Cognitive Semantics Vol.1: Concept Structuring Systems. Cambridge, Mass.: MIT Press, 2000.
[60] M. E. Winston, R. Chaffin, and D. Herrmann, A taxonomy of part - whole relations. Cognitive Science, 11(4), 417–444, 1987.

(Received 2022.12.31; Revised 2023.8.3; Accepted 2023.8.6)