How "empirical" is corpus linguistics?: A meta-analysis using formal concept analysis

Kazuho KAMBARA[1,2], **Yuki Sugawara**[3], **Norihisa TAKAHASHI**[2]
[1] National Institute of Information and Communications Technology / Kyoto, Japan
[2] Ritsumeikan University / Shiga, Japan
[3] Osaka University / Osaka, Japan
Correspondence: kambara-k@nict.go.jp

Abstract

Traditionally, corpus linguistics has been positioned as a field that provides methodologies for observing linguistic phenomena and verifying hypotheses derived from linguistic theories (Fillmore, 1992; Gries, 2010b). McEnery and Brezina (2022) summarised the features of corpus linguistic enquiries as 48 principles and treated hypothesis verification as a central feature. However, this characterisation is still open to discussion. Using formal concept analysis (Ganter and Wille, 1999; Ganter et al., 2005) on the abstracts of *International Journal of Corpus Linguistics*, we show that corpus linguistics is a discipline that often aims at hypothesis generation rather than hypothesis testing.

1 Introduction

Since the dawn of corpus linguistics, its effectiveness in observing linguistic phenomena has been recognised. Along with its advantages, its status as a "theory" has also been discussed (McEnery and Hardie, 2012, 147–164). McEnery and Brezina (2022) elegantly summarised 48 principles as foundations of corpus linguistics and treated hypothesis testing as its central notion. However, its validity remains debatable. In this paper, we argue that corpus linguistics is NOT a field based on hypothesis testing but on hypothesis generation. This result does not diminish the scientificity of corpus linguistics in any way.

This paper is structured as follows: Section 2 overviews the proposals in McEnery and Brezina (2022) and introduces our research question. Section 3 explains the methods and procedures employed in our study. Section 4 reports the results of formal concept analysis and shows that hypothesis testing is not necessarily a central notion in corpus linguistics. Section 5 concludes and overviews some possible developments.

2 Towards a philosophy of corpus linguistics

2.1 Corpus linguistics as theories

This section briefly summarises the long-standing debate on the theoreticality of corpus linguistics and introduces the basic tenets of McEnery and Brezina (2022). The theoreticality of corpus linguistics (i.e., how corpus linguistics connects with (or isolates from) linguistic theories) has been debated (cf. Gries, 2010b; McEnery and Hardie, 2012; Tognini-Bonelli, 2001; Teubert, 2007), leading to the discussion of the scientificity of corpus linguistics. McEnery and Brezina (2022) take the debate on theoreticality to the next level by borrowing the notion of falsifiability (Popper, 1972, 1975). Drawing from Popper's ideas, McEnery and Brezina characterised corpus linguistics as a science of hypothesis testing.

The role of corpus linguistics has been among the topics most discussed by many scholars. A well-known illustration of corpus linguists dates back to Fillmore (1992). In contrast to armchair linguists, corpus linguists are portrayed as someone analysing large datasets to draw quantitative generalisations without paying much attention to theoretical details. However unlikely Fillmore's portraits of corpus linguists are (cf. Gries, 2010b), defining characteristics of corpus linguists has been discussed seriously.

One of the central points in this debate is whether corpus linguistics is a theory or not. Some scholars (cf. Gries, 2010b; McEnery and Hardie, 2012) emphasised the role of corpus linguistics as "tools" of linguistic theories, while others (cf. Tognini-Bonelli, 2001; Teubert, 2007) argued for the theoreticality of corpus linguistics. Although it is highly controversial to assume a set of data and methodology can qualify as "theories", corpus linguistic enquiries often appear theory-independent ones since the authentic data (almost always) "be-

tray" our intuitions, which leads to a more accurate understanding of our authentic language use.

The debates of the theoreticality of corpus linguistics ultimately can lead to the philosophy of science. Philosophy of science deals with ontological and epistemological problems: The former corresponds to the question of "what something is" and the latter to "how we know what something is" (cf. Dennett, 1996; Kambara and Yamanaka, 2023). More specifically, philosophers of science attempt to reveal the kinds of targets of corpus linguistic enquiries (i.e., ontological enquiries) and how corpus linguists know those entities (i.e., epistemological enquiries).

McEnery and Brezina (2022) attempted to construct a full-fledged philosophy of science following the ideas of Popper (1972, 1975). Popper's philosophy of science is perhaps best known for positioning falsifiability as the central notion of scientific enquiries. Falsifiability is the ability to falsify (prove to be false) a hypothesis based on a single observation (McEnery and Brezina, 2022, 42). For instance, someone argued, "Martians' telepathic thought is the strong predictors of distinguishing synonymous pairs (e.g., *sofa* vs. *couch*)". Linguists would not even consider the statement's validity since it cannot be confirmed objectively in any way imaginable. The notion of falsifiability plays a crucial role in deciding which hypothesis is worthy of serious contemplation. Unlike the hypothesis regarding the Martians' interruptions to our daily communications, the hypothesis "A register/genre is a strong predictor of distinguishing synonymous pairs (e.g., *sofa* vs. *couch*)" is much more appealing to corpus linguists since its validity can be examined using various corpus linguistic techniques.

The notion of falsifiability assumes that the heart of scientific enquiries is in verifying hypotheses. Borrowing this idea from the philosophy of science can help us understand the corpus linguistic endeavour more precisely. Due to its theoretical importance, various favourable reviews (Curry, 2023; Levin, 2023; Wu, 2023) have been published, suggesting that many corpus linguists agree with the assumption that corpus linguistics is a science of hypothesis verification.

2.2 Scientificity of corpus linguistics

This section challenges the view that corpus linguistics is a science of hypothesis verification. We argue that this perspective, which centralises hypothesis verification, is problematic because it inherits the limitations of conventional philosophy of science (including Popper's), whose understanding of scientificity heavily depends on the traditional framework of physics. We briefly review current advances in the philosophy of science to argue that positioning hypothesis verification as the central notion of corpus linguistics is debatable.

Philosophers of science in the first half of the 20th century divided scientific inquiry into (i) **the context of justification** and (ii) the **context of discovery** (Reichenbach, 1938). The former was not regarded as a process of scientific inquiry because of psychological factors, and the latter was regarded as the central notion in scientific enquiry. This division was influenced by the position of **logical positivism** in the first half of the 20th century. Logical positivism attempted to create the movement for **Unity of Science** aiming to understand science as a whole employing ideas of mathematics, logic, and physics (Cat, 2024). Logical positivism eliminated the context of discovery, which is an "illogical" process, and emphasised the importance of context of justification in the scientific enquiry of the philosophy of science (Schickore, 2022).

In contrast, philosophers of science in the latter half of the 20th century, the target of the philosophy of science expanded beyond the field of science initially attempted by logical positivism. The philosophy of special sciences (i.e., philosophy of individual scientific fields), which closely examines the case of a specific field, became the central endeavour (Fodor, 1974). This movement was embodied in the 1990s by a movement titled **Disunity of Science** by the Stanford School led by John Dupré, Ian Hacking, Peter Galison, Patrick Suppes, and Nancy Cartwright (Galison and Stump, 1996). The Stanford School rejected logical positivism's attempt to describe a unified world of science and helped to redirect analysis toward describing fragments of individual science. Descriptive science, such as biology, has been adopted as an object of analysis in the philosophy of individual science, which does not necessarily emphasise the context of justification, unlike physics.

For instance, evolutionary biologists are likely to describe the different shapes of beaks in various ecological niches (Skipper and Millstein, 2005). Scientists do not regard these works as irrelevant just because they do not (in a strict sense) verify a hypothesis. Moreover, neuroscientists are more interested in the mechanisms of humans' neural

networks (Machamer et al., 2000; Craver, 2007). Again, these enquiries are "scientific" enough even though they do not aim to verify a hypothesis. These cases suggest that it is not realistic to build scientific foundations for a given field just by borrowing ideas that are originated from the framework of physics.

From a linguistic point of view, it is well-known that methodologies of generative grammar are inspired by those of physics (Harris, 2021, 11). As often discussed, the guiding principles of corpus linguistics are far from those of generative grammar since corpus linguists emphasise the importance of solidifying observational foundations (Leech, 1992). For this reason, it is debatable if the core enterprise of scientific enquiries resides in the verification of hypotheses.

In this paper, we aim to observe the qualitative characteristics of corpus linguistics empirically, which can confirm the generalisation made by McEnery and Brezina. If our discussion is on the right track, hypothesis verification should not be observed often in published corpus linguistic research papers. In this sense, our enquiry can be positioned as a meta-analysis of corpus linguistics.

3 Methods

This section explains the methods used in this study. To observe the characteristics of corpus linguistics, we extracted the abstracts of *International Journal of Corpus Linguistics* (IJCL) and manually annotated them to conduct formal concept analysis (FCA). In the following, after explaining the data extraction procedure, we overview the characteristics of FCA in Section 3.1, introduce the annotation strategies in Section 3.2, and describe the procedure of analysis in Section 3.3.

3.1 Formal Concept Analysis (FCA)

Formal Concept Analysis (FCA) is a method developed by Ganter and Wille (1999). It was developed as a lattice theory in applied mathematics. It deals with qualitative data in the form of $i \times j$. FCA provides a powerful way to visualise the structure of a given data, especially their implicational structures. It has been applied in language studies (Priss, 1998, 2005; Hasebe and Kuroda, 2009; Kuroda, 2015).

For instance, let us say we are interested in the semantic relations among person-denoting nouns (i.e., *person*, *adult*, *child*, *man*, *boy*, *woman*, *girl*). These nouns can be analysed in the form of Table 1. Each semantic feature is coded in a binary fashion. Note that attributes with the value "1" indicate that the noun has the given attributes (i.e., *girl* is YOUNG and FEMALE). Therefore, *person*, the most general term, is coded as 0 for all attributes, representing the absence of specific features (YOUNG, OLD, MALE, FEMALE).

Table 1: A formal concept of person-denoting nouns

	YOUNG	OLD	MALE	FEMALE
person	0	0	0	0
adult	0	1	0	0
child	1	0	0	0
man	0	1	1	0
boy	1	0	1	0
woman	0	1	0	1
girl	1	0	0	1

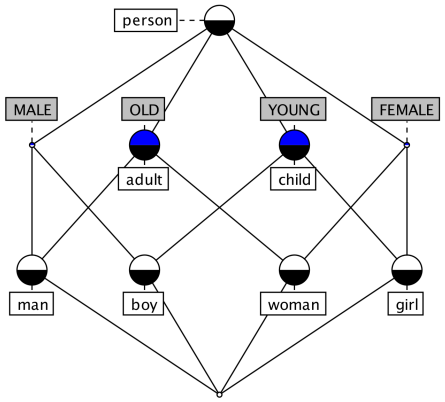

Figure 1: The lattice structure of person-denoting nouns in English

We obtain the lattice in Figure 1 by importing the data to Concept Explorer. This lattice is called concept lattice and represents the class-inclusion relations. White boxes show the names of objects (e.g., *person*, *adult*, *child*, ...), and grey boxes show the names of attributes that classify the given objects. The lattice shows the following implicational relations in (1) from Table 1.

(1) a. The referent of *person* subsumes those of the other six nouns.
b. The nouns *man* and *woman* are special cases of *adult* and of *person* (i.e., The nouns *man* and *woman* are hyponyms of *adult* and *person*).
c. The nouns *boy* and *girl* are special cases of *child* and of *person* (i.e., The nouns boy and girl are hyponyms of *child* and *person*).

d. Unlike the attributes OLD and YOUNG, the attributes MALE and FEMALE do not possess any unique objects, and they are distributed to the special cases of objects possessing OLD or YOUNG.

The structure visualised in Figure 1 is known as a **Hasse diagram**.

Since the conceptual structure of data in Table 1 is relatively straightforward, the interpretations of the concept lattice in Figure 1 are pretty simple. However, identifying and understanding the structure of a large table with a substantial number of rows and columns is labour-intensive. Using FCA can mediate such processes. This paper aims to identify and understand the nature of corpus linguistics using such techniques. The attributes used in this study are explained in Section 3.2.

3.2 Annotation strategies

Because the conceptual structure in a given table is not necessarily evident in advance, the resulting classification may produce a non-optimal lattice. Discarding an object or attribute can help create an optimal concept lattice in FCA. In discarding variables, such actions should be justified by theoretically probable reasons. To achieve this goal, we devised the following attributes to classify the given corpus linguistic research.

(2) Types of goals:
   a. `is_theoretical`: 1 iff the given paper's goal was motivated theoretically, 0 otherwise.
   b. `is_educational`: 1 iff the given paper's goal was motivated educationally, 0 otherwise.
   c. `is_methodological`: 1 iff the given paper's goal was motivated methodologically, 0 otherwise.

(3) Types of "methods":
   a. Types of approach:
      (i) `verifies_hypothesis`: 1 iff the given paper's goal is to verify a hypothesis, 0 otherwise.
      (ii) `presents_hypothesis`: 1 iff the given paper aimed to generate or present a new hypothesis via describing certain phenomena, 0 otherwise.
   b. Kinds of targets:
      (i) `target_is_spoken`: 1 iff the target of analysis was spoken, 0 otherwise.
      (ii) `target_is_written`: 1 iff the target of analysis was written, 0 otherwise.
   c. Characteristics of targets:
      (i) `corpus_is_balanced`: 1 iff the analysed corpus (or its fragments) was balanced in its own right, 0 otherwise.
      (ii) `corpus_is_representative`: 1 iff the analysed corpus (or its fragments) was representative, 0 otherwise.
   d. Originality of targets:
      (i) `introduces_new_dataset`: 1 iff the author(s) of the given paper devised a new dataset, 0 otherwise.
      (ii) `dataset_is_shared`: 1 iff `introduces_new_dataset` is 1, AND the author(s) of the presented paper made the new dataset public, 0 otherwise.

(4) Types of phenomena:
   a. `target_is_micro`: 1 iff the analysed target was a specific expression (e.g., word, phrase, construction), 0 otherwise.
   b. `target_is_cross-linguistics`: 1 iff the analysed target was cross-linguistic, 0 otherwise.
   c. `target_is_variation`: 1 iff the analysed target was a variation of some kind (e.g., genre, gender, place), 0 otherwise.

(5) The year of publication:
   a. `is_in_90s`: 1 iff the given paper was published in the 1990s, 0 otherwise.
   b. `is_in_00s`: 1 iff the given paper was published in the 2000s, 0 otherwise.
   c. `is_in_10s`: 1 iff the given paper was published in the 2010s, 0 otherwise.
   d. `is_in_20s`: 1 iff the given paper was published in the 2020s, 0 otherwise.

Though some attributes may seem redundant, the finalised design is intentional. For instance, as for Types of approach, we intentionally devised both `verifies_hypothesis` and

presents_hypothesis to code the purpose of a given paper that verifies and presents a new hypothesis at the same time. Kinds of targets have four possible combinations, as shown in Table 2. Characteristics of targets are certainly nuanced. The attribute corpus_is_balanced is evaluated on whether the author(s) employed a balanced corpus. Representativeness of corpora is evaluated on how exhaustive the authors collected the given data. For instance, if an author decided to analyse the language use in e-mail exchanges and collect only a handful, corpus_is_balanced is coded as 0.

3.3 Procedures

We first extracted abstracts of all articles published in *International Journal of Corpus Linguistics* (IJCL) from 1996 to 2023. We excluded a total of 53 book reviews and contributions in special issues, as the abstracts of these articles could not be retrieved automatically. As a result, we chose 440 articles for exhaustive qualitative analysis.

For each of the 440 articles, we manually and semi-automatically annotated the features introduced in Section 3.2. After standardising the article names, we used the final file to input Concept Explorer 1.3 for formal concept analysis. However, using the whole data significantly slowed the program's execution, so we randomly sampled 50 articles for formal concept analysis. All 440 annotated abstracts are available on Open Science Framework (OSF).

4 Analysis

4.1 Classification without optimisation

This section reports the results of FCA and their interpretations. As explained, the "uncompromised" lattice can yield non-optimal classification, as shown in Figure 2 with red colliding lines, which is typical when many attributes are included. To arrive at an optimal solution, we can discard either (i) some objects (for possible misclassifications) or (ii) some attributes. For the purpose of achieving a more readable and interpretable lattice, and under the assumption that the object classification is sound, we proceeded by selectively removing attributes introduced in Section 3.2.

4.1.1 Classification based on the goals

Removing all the attributes other than the types of goals produces a simplified lattice as in Figure 3. It shows that (i) the attributes is_theoretical, is_methodological, and is_educational each possess unique generating objects (meaning each goal type is the sole characteristic of at least one research paper), and (ii) no articles were classified as theoretical and educational, while the mixture of methodological motivations with theoretical or educational motivations was observed. As the node in the lattice shows, the most frequent motivations are methodological ($38\% = 19/50$), which aligns with the perspective that corpus linguistic enquiries are often viewed as "tools" for theories.

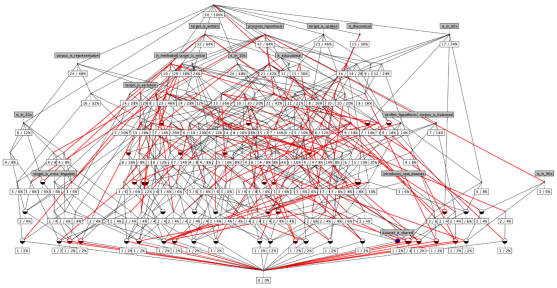

Figure 2: The "uncompromised" lattice of sampled articles using all attributes (with the proportions instead of article ids)

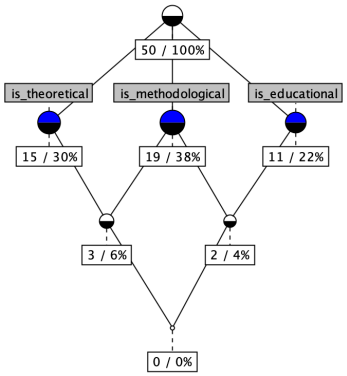

Figure 3: The concept lattice of goals (with the proportions instead of article IDs)

4.1.2 Classification based on the "methods"

Since the number of attributes related to "methods" is quite large, the classification lattice becomes more complex than the other types. Figure 4 is the lattice based on the Types of "methods", in which only two clear implications are read:

(6) a. If a given research paper's data set is shared (in the sense of dataset_is_shared), it introduces a new data set (i.e., corpora) which is a collection of written language (target_is_written), and it

Table 2: Possible combinations of `target_is_spoken` and `target_is_written` and their examples

	`target_is_spoken` = 1	`target_is_spoken` = 0
`target_is_written` = 1	A corpus of written and spoken language	A corpus of written language
`target_is_written` = 0	A corpus of spoken language	A corpus of other language (e.g., sign language), or NA

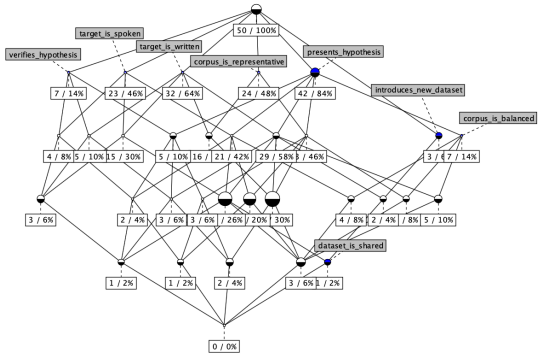

Figure 4: The concept lattice of "methods" (with the proportions instead of article IDs)

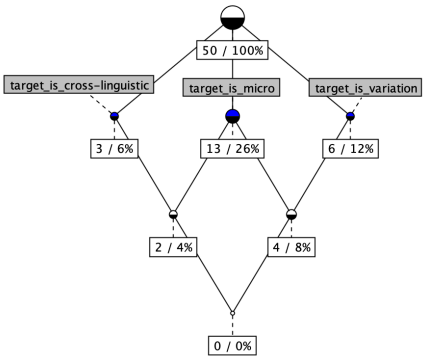

Figure 5: The concept lattice of phenomenon types (with the proportions instead of article IDs)

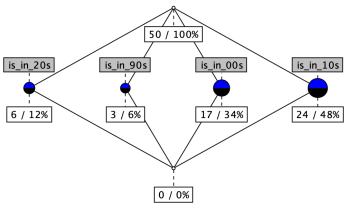

Figure 6: The concept lattice of publication year (with the proportions instead of article IDs)

presents a hypothesis

b. If a given research paper utilises a balanced corpus (i.e., `corpus_is_balanced`), it presents a hypothesis.

Though the first implication is interesting enough, constructing a corpus of written language can be due to the effect of random sampling. However, it is likely for corpus linguists to construct a corpus of written language since it is more accessible than spoken ones. In addition, the latter parts of both implications arrive at presentations of hypotheses, suggesting that verifying hypotheses in corpus linguistic enquiries is not as central as conventionally assumed.

4.1.3 Classification based on types of phenomena

Similar to the classification lattice in Figure 3, the classification lattice based on the types of phenomena is easy to understand. Figure 5 is the lattice using only the phenomenon types for its attributes. As can be read from the lattice, `target_is_cross-linguistic` and `target_is_variation` are mutually exclusive. Since most of the variation research focuses on the distributions of particular expression(s), it is technically challenging to combine cross-linguistic enquiries with variation research. However, if we ignore such variations, we can compare the corresponding expressions cross-linguistically. The lattice shows some of the practical constraints in a corpus linguistic research.

4.1.4 Classification based on the years

Figure 6 shows that the concept lattice based on publication periods is not as "interesting" as the others. This is because all attributes are mutually exclusive, and it only shows the proportions of each time period. It shows that most of the investigated articles were published in the 2010s, reflecting our extraction procedures' limitations. As can be read from the lattice, the raw frequency of `is_in_20s` is larger than that of `is_in_90s`, suggesting that the frequency of publication is accelerated, considering our data only contains articles published in recent years.

4.2 Hypotheses testing in corpus linguistics

As the concept lattice in Section 4.1.2 shows, the role of hypothesis verification is less central than conventionally assumed. This tendency is suggested by the fact that the node labelled `verifies_hypothesis` in Figure 4 does not solely define a unique concept and accounts for only 14 articles (14%). Instead, the lattice strongly shows that presentations of hypotheses are more dominant in the given dataset since the attribute presents_hypotheses possesses unique objects and has 42 unique objects (82%).

The characterisation of corpus linguistics by McEnery and Brezina (2022) relies heavily on the characterisation of scientific enquiries by Popper (1972, 1975). As already pointed out in Section 2.2, the centrality of hypothesis verification in corpus linguistics can be debatable. The simple summarisation of attributes suggests that hypothesis generation (as captured by the `presents_hypothesis` attribute) is more widespread than hypothesis verification.

However, this fact does not diminish the scientificity of corpus linguistics. Not all scientific fields aim to verify hypotheses; some simply emphasise the importance of describing the nature of a target in interest. In some subfields of biology, analysts do not always have an overall understanding of investigated creatures, which usually motivates their empirical enquiries (Kampourakis and Uller, 2020). Like these biologists, corpus linguists often begin without knowing precisely how a given expression behaves in a specific discourse. Instead, they usually devise a systematic procedure for observing various instances of authentic language use. If linguists could know the "inside" of corpus data, encountering unexpected instances becomes impossible. Fillmore (1992, 35) pointed out that corpus data allows linguists to observe data without unnecessary biases.

The descriptive tendency of corpus linguistics invites gap-spotting approaches (Alvesson and Sandberg, 2013), in which researchers identify the "gap" in previous studies to construct their research questions[1]. As discussed, corpus linguists do not know the contents of the investigated data, which easily allows them to create a research question. For instance, if a researcher finds a frequently discussed topic in theoretical or applied linguistics, she can ask how actual speakers realise such a phenomenon. For instance, Gries (2006) discussed the polysemous network of the verb *run*. A lexical item's polysemy network has been discussed in cognitive linguistic literature (Lakoff, 1987). However, how such a network is structured from attested cases had not been clarified. Gries identified senses of the verb *run* and demonstrated how corpus-linguistic techniques contribute to identifying the quantitative and qualitative aspects of polysemous words. This work is a typical case of the gap-spotting approach because linguists did not know the behaviour of the word *run*.

However, as repeatedly emphasised, the descriptive nature of corpus linguistics does not diminish the scientificity of corpus linguistics. Instead, it suggests that corpus linguistic enquiries should be seen as descriptive science like biology (or maybe even ecology). These fields of descriptive science contribute to a realistic understanding of entities in the real world. Since corpus linguists have emphasised the importance of observing attested cases, it is more natural to assume that corpus linguistics is a science of discovery rather than a science of verification.

4.3 Corpus linguistics as a "method" (all over again)

In qualitative analysis using FCA, analysts must carefully select the appropriate attributes that represent some significant characteristics of the target. As discussed, uncompromised classification results in non-optimal resolution (See Figure 2). Based on the discussion that hypothesis verification is not the central notion in corpus linguistics, we selected four attributes: (i) `is_theoretical`, (ii) `is_methodological`, (iii) `presents_hypothesis`, and (iv) `target_is_micro`. As Figure 7 results in an optimal classification, these attributes can represent a typical research project in corpus linguistics.

Figure 7 shows that all research projects are classified into three cross-cutting categories by the above-mentioned attributes `is_theoretical`, `presents_hypothesis`, and `is_methodological`. Among these major attributes, the attribute `presents_hypothesis` is the most widespread (84%), and some projects are purely theoretical (1 unique object) or

[1] Alvesson and Sandberg (2013) criticise the overuse of gap-spotting approaches in social science because such approaches do not invite a novel researches. We refrain from stating that it is preferable or non-preferable for corpus linguists to follow such practices.

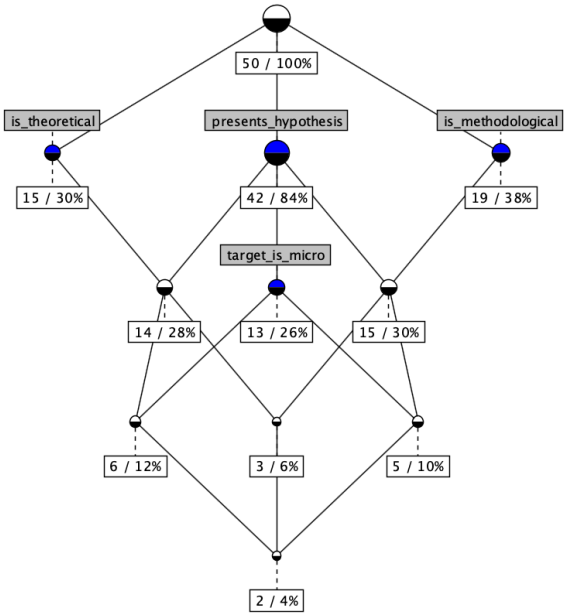

Figure 7: The concept lattice of goals and their approaches

methodological (4 unique objects).

One widely acknowledged advantage of using corpora is that they offer an objective method to observe language use (Fillmore, 1992). The lattice in Figure 7 seems to support the conception of "corpus linguistics as a methodology" because most research projects aim to present a novel hypothesis based on observation of attested data. Some scholars argue that corpus linguistics is a theory of its own, while others treat it as a methodology (McEnery and Brezina, 2022, 147–162). Our analysis empirically suggests that the latter characterisation (corpus linguistics as a methodology/science of discovery) is more fitting than the former.

We repeatedly emphasised that observation of attested data and developing observational tools are central to corpus linguistics, which accords with the statement that hypothesis presentation/generation is more pervasive than hypothesis verification. This tendency was also confirmed in (Gilquin and Gries, 2009). The fact that corpus linguists do not verify hypotheses informed by linguistic theories as often as expected does NOT diminish the scientificity of corpus linguistics. Developments of observational tools cannot be separated from the developments of science. For instance, scientists could not have arrived at a better understanding of creatures' microstructures without the help of electron microscopes. Likewise, without the help of corpora, linguists could have never understood how our intuitions are "betrayed" by attested data.

However, the notion of corpus-as-method does not entirely depend on a specific linguistic theory. In developing concordance tools and constructing a new set of corpora, corpus linguists borrow various notions from neighbouring fields (e.g., Natural Language Processing; NLP). In analysing the given set, analysts exploratorily annotate the given data (cf. Gries, 2010a; Kambara et al., 2023) to see if any combinations of the given variable significantly contribute to the analysis of the given phenomenon. These practices are not deductively derived from the predictions of linguistic theories. Instead, they embrace the "irregularities" found in the data, which accords with the pragmatistic conceptions of science (cf. Quine, 1960, 1961). The pragmatic conceptions of science refer to the gradual progress of scientific knowledge employing all the available resources (Kambara and Yamanaka, 2023; Nefdt, 2023).

In this context, it can be said that the central role of corpus linguistics is to discover **real patterns** hidden in the data (Dennett, 1991). This is the task of systematically capturing the complexity and diversity of actual language use, something that theorists can overlook. If one of the important goals of linguistics is to understand the complex phenomenon of language, then the inventories of corpus linguistic techniques to discover patterns provides an indispensable contribution to the entire field of linguistics. Corpus linguists as discoverers of real patterns can provide a more sophisticated understanding to the debates on corpus-as-method.

5 Conclusion

In this paper, we empirically analysed abstracts published in *International Journal of Corpus Linguistics* (IJCL) and applied Formal Concept Analysis (FCA) to gain a deeper understanding of the field's characteristics. The FCA results strongly suggest that corpus linguistics operates as a science of discovery (akin to descriptive fields like biology) rather than fundamentally as a science of verification. While these findings partially align with the "corpus-as-method" perspective, we argue that this descriptive, discovery-oriented nature necessitates recognising two vital points: (i) corpus linguistics constitutes a distinct scientific field, and (ii) the development of observational tools and procedures is central to its scientific endeavour.

Limitations

Two issues remain unsolved.

First, as previously noted in Section 3.2, relying solely on abstracts for annotation risks distorting the authors' actual intentions and the full scope of their research. Future work should develop a more comprehensive strategy, such as analyzing the full-text content, to address this limitation.

Secondly, our study focused on the qualitative conceptual structure of the papers, rather than conducting a broad quantitative analysis. For our characterisation of corpus linguistics to be fully robust, future studies should aim to quantitatively replicate similar tendencies across a wider population of articles, including those published in related journals.

Acknowledgement

We thank the comments from the three anonymous reviewers. The earlier version of this paper was read at THE JAECS CONFERENCE 2023. Portions of this work were supported by JSPS KAKENHI Grant Number 24K16143. All remaining errors are ours.

References

Mats Alvesson and Jörgen Sandberg. 2013. *Constructing Research Questions: Doing Interesting Research*. Sage Publications, London.

Jordi Cat. 2024. The unity of science. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Spring 2024 edition. Metaphysics Research Lab, Stanford University.

Carl F. Craver. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press, Oxford.

Niall Curry. 2023. Review of McEnery & Brezina (2022): *Fundamental Principles of Corpus Linguistics*. *International Journal of Corpus Linguistics*, 28(2):278–283.

Daniel C Dennett. 1991. Real patterns. *Journal of Philosophy*, 88(1):27–51.

Daniel C. Dennett. 1996. *Kinds of Minds: Toward an Understanding of Consciousness*. Basic Books, New York.

Charles J Fillmore. 1992. "Corpus linguistics" vs. "computer-aided armchair linguistics". In Jan Svartvik, editor, *Directions in Corpus Linguistics: Proceedings from a 1991 Nobel Symposium on Corpus Linguistics*, pages 35–66. Mouton de Gruyter, Berlin.

Jerry A. Fodor. 1974. Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, 28(2):97–115.

Peter Galison and David J. Stump, editors. 1996. *The Disunity of Science: Boundaries, Contexts, and Power*. Stanford University Press, Stanford.

Bernhard Ganter, Gerd Stumme, and Rudolf Wille, editors. 2005. *Formal Concept Analysis: Foundations and Applications*. Springer, Dresden.

Bernhard Ganter and Rudolf Wille. 1999. *Formal Concept Analysis: Mathematical Foundations*. Springer, Berlin, Heidelberg.

Gaëtanelle Gilquin and Stefan Th. Gries. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, 5(1):1–26.

Stefan Th. Gries. 2006. Corpus-based methods and cognitive semantics: The many senses of *to run*. In Stefan Th. Gries and Anatol Stefanowitsch, editors, *Corpora in Cognitive Linguistics: Corpus-Based Approach to Syntax and Lexis*, pages 57–99. Mouton de Gruyter, Berlin.

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

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

Randy Allen Harris. 2021. *The Linguistics Wars: Chomsky, Lakoff, and the Battle over Deep Structure*. Oxford University Press, Oxford.

Yoichiro Hasebe and Kow Kuroda. 2009. Extraction of English ditransitive constructions using formal concept analysis. In *23rd Pacific Asia Conference on Language, Information and Computation*, pages 678–685.

Kazuho Kambara, Hajime Nozawa, and Takeshi Takahashi. 2023. Differentiating valence patterns: A quantitative analysis based on formal and semantic attributes. *Constructions*, 15(2).

Kazuho Kambara and Tsukasa Yamanaka. 2023. Philosophy of data science for corpus linguistics: A pragmatistic point of view. *Annals of the Japan Association for Philosophy of Science*, 32:47–73.

Kostas Kampourakis and Tobias Uller, editors. 2020. *Philosophy of Science for Biologists*. Cambridge University Press, Cambridge.

Kow Kuroda. 2015. Formal concept analysis meets grammar typology. In *Proceedings of the Twenty-first Annual Meeting of the Association for Natural Language Processing*, pages 329–332.

George Lakoff. 1987. *Woman, Fire and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press, Chicago.

Geoffrey Leech. 1992. Corpora and theories of linguistic performance. In Jan Svartvik, editor, *Directions in Corpus Linguistics: Proceedings from a 1991 Nobel Symposium on Corpus Linguistics*, pages 105–122. Mouton de Gruyter, Berlin.

Magnus Levin. 2023. Tony McEnery and Vaclav Brezina. Fundamental principles of corpus linguistics. Cambridge: Cambridge University Press, 2022. 313 pp. ISBN 978-1-1071-1062-5. *ICAME Journal*, 47(1):141–143.

Peter Machamer, Lindley Darden, and Carl F. Craver. 2000. Thinking about mechanisms. *Philosophy of Science*, 67(1):1–25.

Tony McEnery and Vaclav Brezina. 2022. *Fundamental Principles of Corpus Linguistics*. Cambridge University Press, Cambridge.

Tony McEnery and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, Cambridge.

Ryan M. Nefdt. 2023. *Language, Science, and Structure: A Journey into the Philosophy of Linguistics*. Oxford University Press, Oxford.

Ryan M. Nefdt. 2024. *The Philosophy of Theoretical Linguistics: A Contemporary Outlook*. Cambridge University Press, Cambridge.

Barbara H. Partee, Alice Ter Meulen, and Robert E. Wall. 1987. *Mathematical Methods in Linguistics*. Kluwer Academic Publishers, Dresden.

Karl R. Popper. 1972. *Objective Knowledge: An Evolutionary Approach*, revised edition. Oxford University Press, Oxford.

Karl R. Popper. 1975. *Conjectures and Refutations: The Growth of Scientific Knowledge*, 5 edition. Routledge, London.

Uta Priss. 1998. The formalization of WordNet by methods of relational concept analysis. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 179–196. MIT Press, Cambridge, Mass.

Uta Priss. 2005. Linguistic applications of formal concept analysis. In Bernhard Ganter, Gerd Stumme, and Rudolf Wille, editors, *Formal Concept Analysis: Foundations and Applications*, pages 149–160. Springer, Dresden.

Willard Van orman Quine. 1960. *Word and Object*. MIT Press, Cambridge, Mass.

Willard Van orman Quine. 1961. *From a Logical Point of View: 9 Logico-Philosophical Essays*. Harper & Row Publishers, New York.

Hans Reichenbach. 1938. *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*. University of Chicago Press, Chicago.

Jutta Schickore. 2022. Scientific discovery. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Winter 2022 edition. Metaphysics Research Lab, Stanford University.

Robert A. Skipper and Roberta L. Millstein. 2005. Thinking about evolutionary mechanisms: Natural selection. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2):327–347.

Anna Teubert, Wolfgang Cermakova. 2007. *Corpus Linguistics: A Short Introduction*. Continuum International Publishing Group Ltd., London.

Elena Tognini-Bonelli. 2001. *Corpus Linguistics at Work*. John Benjamins, Amsterdam.

Chenghui Wu. 2023. Book review: *The Fundamental Principles of Corpus Linguistics*. Tony McEnery and Vaclav Brezina. *Digital Scholarship in the Humanities Digital Scholarship in the Humanities*, 38(2):916–917.

A A mathematical characterisation of lattice structures

This appendix provides a formal mathematical characterisation of the lattice structures and concepts (such as the Hasse diagram) used in this paper's Formal Concept Analysis (FCA). Our presentation focuses on illustrating the basic properties of lattice structures and is simplified for explanatory purposes, thus differing slightly from the more formal treatment found in the original characterisation by Ganter and Wille (1999). See Partee et al. (1987) for an introductory explanation of related mathematical concepts.

Definition 1 (partial order and partially ordered set)**.** A binary relation R on a set X satisfies the following conditions 1, 2 and 3.

1. $\forall\, x \in X,\ x\,\mathrm{R}\,x$
2. $\forall\, x, y, z \in X,\ (x\,\mathrm{R}\,y \wedge y\,\mathrm{R}\,z) \Longrightarrow x\,\mathrm{R}\,z$
3. $\forall\, x, y \in X,\ (x\,\mathrm{R}\,y \wedge y\,\mathrm{R}\,x) \Longrightarrow x = y$

Then, the relation R is referred to as **partial order**, the pair (X, R) as a **partially ordered set**.

Definition 2 (lattice)**.** Let $(X, \preceq)$ be a non-empty finite partially ordered set. If $(X, \preceq)$ satisfies following conditions 1 and 2, then $(X, \preceq)$ is called a **lattice**.

1. $\forall x, y \in X$, $\exists z \in X$ such that z satisfies the conditions (a) and (b).
   (a) $z \preceq x \wedge z \preceq y$
   (b) $\forall w \in X$, $w \preceq x \wedge w \preceq y \Longrightarrow w \preceq z$
2. $\forall x, y \in X$, $\exists z \in X$ such that z satisfies the conditions (a) and (b).
   (a) $x \preceq z \wedge y \preceq z$
   (b) $\forall w \in X$, $x \preceq w \wedge y \preceq w \Longrightarrow z \preceq w$

Example 3. Let X be a non-empty finite set and $\preceq$ the binary relation on the powerset $\mathfrak{P}(X)$ defined, for $A, B \in \mathfrak{P}(X)$, by $A \preceq B \overset{\text{def}}{\iff} B \subseteq A$. Then $\preceq$ is a partial order on $\mathfrak{P}(X)$. Thus, for any subset Y of $\mathfrak{P}(X)$, $(Y, \preceq)$ is a partially ordered set. Moreover, for any $A, B \in Y$, we assume that $A \cup B \in Y$ and $A \cap B \in Y$ hold. Then $A \cup B$ and $A \cap B$ satisfy the conditions of the definition 2 with respect to $\preceq$. Therefore $(\mathfrak{P}(X), \preceq)$ is a lattice.

Definition 4 (cover relation)**.** Let $(X, \preceq)$ be a partially ordered set. For $x, y \in X$ with $x \prec y$ (that is, $x \preceq y$ and $x \neq y$), we say that y **covers** x if there is no $z \in X$ such that $x \prec z \prec y$.

Definition 5 (Hasse diagram)**.** Let $(X, \preceq)$ be a non-empty finite partially ordered set. If a graph satisfies the following three conditions, then the graph is called the **Hasse diagram** of $(X, \preceq)$:

(i) The vertex set is X.

(ii) If $x \prec y$ holds, then the vertex y is positioned above the vertex x.

(iii) If y covers x, then give the edge from y to x.

Example 6. Consider the set $X = \{1, 2, 3, 4, 6, 12\}$ of positive divisors of 12. We define a partial order $\preceq$ on X by divisibility: for $x, y \in X$, $x \preceq y \overset{\text{def}}{\iff} x$ divides y.

The cover relations in this partially ordered set $(X, \preceq)$ are

$$1 \prec 2,\ 1 \prec 3,\ 2 \prec 4,\ 2 \prec 6,\ 3 \prec 6,\ 4 \prec 12,\ 6 \prec 12.$$

The corresponding Hasse diagram is given in Figure 8.

It is easy to see that $(X, \preceq)$ forms a lattice, since any two elements of X admit a greatest common divisor and a least common multiple within X.

Example 7 (Application to linguistic categorisation)**.** As shown in Table 1, some person-denoting nouns can be classified using the two properties, (i) age (Old/Young) and gender (Male/Female). The subsets of four attributes $\{$Young, Old, Male, Female$\}$ can represent the semantics of each noun.

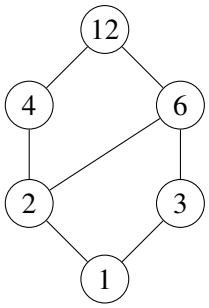

Figure 8: Hasse diagram of the divisors of 12 ordered by divisibility

$$\text{person} = \emptyset,$$

$$\text{adult} = \{\text{Old}\},$$

$$\text{child} = \{\text{Young}\},$$

$$\text{man} = \{\text{Old, Male}\},$$

$$\text{boy} = \{\text{Young, Male}\},$$

$$\text{woman} = \{\text{Old, Female}\},$$

$$\text{girl} = \{\text{Young, Female}\}.$$

Now, for the finite set $X = \{$Male, Female, Young, Old$\}$, if we define the subset Y of $\mathfrak{P}(X)$ by

$$Y = \left\{ \begin{matrix} \text{person}, X, \{\text{Male}\}, \{\text{Female}\}, \\ \text{adult, child, man, boy, woman, girl} \end{matrix} \right\},$$

and the binary relation $\preceq$ on Y by the similar manner in Example 3, then $(Y, \preceq)$ is a lattice.

Therefore, the term "person" admits a lattice structure, which can be visualised using the Hasse diagram as shown in Figure 1.