Modeling and Corpus Methods in Experimental Philosophy

Louis Chartrand

March 13, 2022

Abstract

Research in experimental philosophy has increasingly been turning to corpus methods to produce evidence for empirical claims, as they open up new possibilities for testing linguistic claims or studying concepts across time and cultures. The present article reviews the quasi-experimental studies that have been done using textual data from corpora in philosophy, with an eye for the modeling and experimental design that enable statistical inference. I find that most studies forego comparisons that could control for confounds, and that only a little less than half employ statistical testing methods to control for chance results. Furthermore, at least some researchers make modeling decisions that either do not take into account the nature of corpora and of the word-concept relationship, or undermine the experiment's capacity to answer research questions. I suggest that corpus methods could both provide more powerful evidence and gain more mainstream acceptance by improving their modeling practices.

Keywords: experimental philosophy, corpus methods, statistical inference, models

1 Introduction

In the humanities and social science, corpora (singular: corpus) are datasets composed of textual data that have been put together specifically for research, typically with a certain type of research question in mind. Methods employed to analyze them in the pursuit of a research program usually involve some computer assistance, both to address the size of these datasets, which often makes close reading of the entire collection impossible, and to allow for the use of statistical and algorithmic tools that would be cumbersome to compute on a naked brain. Depending on the discipline, corpus analysis approaches and methods are often found under labels such as "digital humanities", "corpus linguistics", "computational linguistics" or "computer-assisted text analysis".

Philosophy has been relatively slow to adopt corpus methods. Nevertheless, while the phenomenon has remained marginal, philosophers have been interrogating texts with computers since the 1970s, as witnessed by Meunier et al.'s System for Text and Content Analysis (1976), or McKinnon's statistical profile of Kierkegaard's works (1970). Traditionally, text analysis was conceived mostly as a way to assist reading and interpretation, be it by developing algorithms to discover patterns that close reading would miss (Forest and Meunier 2000; Meunier, Forest, and Biskri 2005; Sainte-Marie et al. 2011; Danis 2012) or by using computational resources to exploit massive corpora (Chartier et al. 2008; Malaterre and Chartier 2019). In these studies, the research object is contained within the corpus itself, such that there is no need to generalize any conclusion beyond the data available—for instance, in Alfano (2018), the object is the concepts of *drive*, *instinct* and *virtue* in Nietzsche's works, and the corpus just *is* Nietzsche's monographic production, so whatever conclusions Alfano draws about those concepts, they concern text that he can observe and describe. Thus, the methods developed for this early corpus-based philosophy were mostly descriptive: descriptive statistics like word counts and co-occurrence statistics (tells which pairs of words are often found together), unsupervised machine learning algorithms like topics models (models that identify discursive topics in a corpus), clustering algorithms (that find clusters of similar textual objects) or dimensionality reduction techniques (that express textual objects through a limited number of semantic dimensions), and visualization techniques.

In other disciplines (Gilquin and Gries 2009; Baroni and Evert 2009), researchers have been using corpora for justification or validation purposes. For instance, one might want to test whether people use visual perception verbs such as "look" and "appear" for their doxastic or their perceptual meaning (Fischer, Engelhardt, and Herbelot 2015) by looking at co-occurrence patterns, or use an corpus of ancient chinese texts to settle disputes about the meaning of a concept among the circles to which those produced those texts (Slingerland and Chudek 2011). Testing hypotheses on corpora can be especially challenging, if only because textual corpora tend to be particularly messy and word occurrences follow unusual distributions. Nevertheless, amid the rise of experimental philosophy, we have seen the appearance of a new paradigm that mobilizes textual corpora not as a way to describe text in new ways, but as a way to understand linguistic behavior beyond the text itself, such as how people use concepts and other thinking resources. Here, linguistic behavior is not something that belongs to, say, an author or a set of authors, but a skill that belongs to a whole community of speakers. Thus, the idea is to take a corpus deemed "representative" of the target population, such as the Corpus of Contemporary American English (COCA) or the British National Corpus (BNC), and, through descriptive methods, human annotations, or other means, provide evidence of certain linguistic behaviors in the target population—one that usually sheds light on an account of a concept or other conceptual device.

Bluhm (2013) connects this new paradigm to experimental philosophy, as it would seem to provide new observables and alternative empirical methods to the classical psychology-inspired vignettes methodology. This call has been echoed in several programmatic papers (Andow 2016; Chartrand 2017; Ulatowski et al. 2020; Caton 2020) and response from the X-Phi community has been mostly positive, as witnessed by the recent *Methodological Advances in Experimental Philosophy* (2019), where corpus methods take up half of the space, and the 2020 *Corpus Fortnight* conference organized by the AXΦ group[1].

Corpus methods open up many horizons for experimental philosophy. Bluhm (2013) suggests that corpora could be used to test hypotheses within an Ordinary Language Philosophy framework. Slingerland and Chudek (2011) suggest that they might tell us "*something* about human cognition in cultures far removed in time" (their emphasis). Overton (2013) sees text mining as a way to "enhance traditional conceptual analysis" through the study of word usage. And so on.

As such, corpus methods are often framed as emerging methods, with a potential unbounded by the test of time. The flip side of this disruptive potential is that their significance, role, and reliability have yet to be established. This is why even in recent papers (e.g. Pease, Aberdein, and Martin 2019; Sytsma et al. 2019; Hinton 2020), philosophers still frame the exploration of corpus methods as a central contribution of their work, as if their contribution was still being assessed.

So far, these assessments have mostly taken the form of proofs of concept, where the researchers apply a method in order to show its potential. However, little has been done to establish the evidential value of corpus-based studies—and one might argue that this could be an obstacle to the widespread adoption of corpus methods. For instance, philosophers could be ambivalent towards the evidential value of corpus studies not because they believe their observations to be flawed, but rather that, from these observations, they cannot confidently draw a conclusion about the theory they had set out to corroborate. Scientists usually address this issue through a set of models and hypotheses that logically connect theories with experimental and statistical statements (Hitchcock 2020). Therefore, in this essay, I wish to review the recent work that has been done in corpus-based (quasi-)experimental[2] philosophy through the lens of modeling practices for generalization and inference.

Drawing from this review, I will argue that experimental philosophers have yet to fully integrate the best research practices that make for good quasi-experimental design. In particular, I will argue that the work under review could be made more convincing and provide stronger evidence by adopting some of the best modeling practices for formulating experimental hypotheses.

[1] Cf. http://axphi.org/corpus-week.

[2] Corpus studies are not properly experimental, as they rely on data that /have been produced beforehand, which makes it much hard to control for possible confounds and ensure variable independence. However, with careful sampling decisions, proper modeling, and a good understanding of the source material, corpora can be leveraged for quasi-experiments. In quasi-experiments, inference from the data to the experimental hypothesis is not licensed by the assurances of the experimental setup, but by reasonable assumptions about the corpus.

In section 2, I present a theory of experimental modeling as it is relevant to the design of corpus quasi-experiments, and how it translates into the design of comparisons, null hypothesis modeling, and the measurement of uncertainty. In section 3, I present a short review of corpus-based quasi-experimental philosophy studies, and how they address comparisons, null hypothesis rejection, and modeling choices. In section 4, I illustrate with examples how proper modeling would have enhanced some of the studies under review. In section 5, I discuss the implications of my findings and propose practices that could improve the value and impact of corpus studies in experimental philosophy.

2 Experimental Modeling and Design for Corpus Studies

Scientific theories, in the formulations that we use for explaining, are typically not empirical statements about specific observations and do not lend themselves to experimentation. This is why scientists usually derive implications from them, which they turn into precise statements about perceptible events that can be used to design an experiment[3].

Take Knobe's (2003) study, which was the first attempt to test the side-effect effect (table 1). In this paper, Knobe presents himself as trying to elucidate the concept of intentional action[4] (first row), and posits that, in the case of side-effects, it might be influenced by valence—which is where you find the substantive theory, $P1$ (*The concept of intentional action is sensitive to valence when it comes to side-effects*). From $P1$, we can derive implications, among which is $P2$ (*Ascriptions of intentionality for side-effects are sensitive to valence*). But $P2$ belongs to the realm of conceptual models (Meunier 2017); for the purpose of experimentation, it still does not concern specific observable events. Thus, "ascriptions of intentionality" is translated into a verbal behavior, and "valence" is translated into "being helpful" or "harmful", giving us $P3$ (*People are more likely to say a side-effect was produced intentionally if it is harmful than if it is helpful*). The experiment is then designed to reenact the situation described in $P3$—which, in experimental philosophy, usually means submitting a vignette to participants—and the statistical hypotheses h_1 and h_0 are the translations of $P3$ and $\neg P3$ into hypotheses about the results of this experiment.

As we go from $P1$ through to $P3$ and h_1, hypotheses are made to be increasingly concrete and precise, so as to eventually be empirically testable. However, this exercise is pointless if testing the experimental hypothesis fails to tell us anything about the substantive theory: therefore, the truth of the hypotheses down the line must somehow warrant the truth of the previous hypotheses. This is accomplished by relying on explicit and implicit background assumptions. For

[3] Cf. Chow (1998) for a more detailed breakdown and explanation of this process.

[4] Since then, Knobe has shifted from attributing the side-effect effect to the concept of intentional action to attributing it to moral judgments affecting human cognition in general (Pettit and Knobe 2009).

example, it is not at all clear that P_3 necessarily implies P_2[5]: it could very well be that something else is getting participants to view harmful side effects as more intentional, like a desire to find a target for blame, or a sense that people who cause harmful effects are norm-violators and thus should be held more responsible of their actions[6]. Thus, there is here a background assumption that P_3 is not due to uncontrolled factors that do not belong to the ascription of intentionality[7]. As such, the breakdown of hypotheses that bridge the gap from substantive theory to experimental and statistical hypotheses not only renders explicit the link between them, but also gives us the conditions under which the evidence can truthfully be said to support the main claim.

Corpus quasi-experiments come with their own challenges when it comes to generalizing from empirical results. Firstly, they are not afforded the convenience of introducing independent variables through randomized interventions, nor can the response encoded in the dependent variable be isolated through an experimental apparatus in a way that facilitates its observation. Therefore, when designing their study, researchers have to model in bottom-up manner, first establishing the possibilities of the corpus and then determining how those possibilities can provide a relevant answer to their research interest. Secondly, in this sort of practice, experimental design cannot be used to shield results from the interaction with confounding factors. Thirdly, perhaps because language is a complex phenomenon where many social, political, cognitive, and pragmatic dynamics are at play, textual corpora tend to be especially noisy, and it is often difficult to distinguish the signal relevant to the study's variables from that noise.

Corpus practitioners have at their disposal a certain number of tools to address these challenges. Firstly, statistical tests themselves can serve to distinguish a true signal from the noise—so long as they are tailored to the task[8]. While most of the statistical tools available expect variables to follow something like a Gaussian distribution, the occurrence of words and other linguistic devices in textual data often follow power laws, so corpus practioners need to ensure that their evaluation methods are appropriate to the phenomena they are observing. Secondly, it is always possible to make use of comparisons. A common type of comparison involves repeating the quasi-experiment on one or more corpora: when a confounding factor is more present in a corpus than the other, it can enable its discovery and control, and when none is found and the effect is observable across corpora, one comes out with a stronger case for its generalization. Comparisons are also used to contrast variables. Most often, the purpose is

[5] I wish to thank an anonymous reviewer for highlighting this.

[6] Cova (2016) lists several of the most prominent explanations that have been proposed to the side effect effect.

[7] This assumption falls into what Lakatos (1970) calls the *ceteris paribus* clause. Given that we can never control for all potential variables, *ceteris paribus* clauses are an inescapable reality of experimentation.

[8] It is worth mentioning there is an ongoing debate on whether statistical testing is desirable in corpus studies (Kilgarriff 2005; Gries 2005; Koplenig 2019). However, those who argue against it also argue against generalizing from results obtained from a corpus.

simply to provide a control condition for the independent variables. However, in the context of corpora, this sort of comparison often plays double duty to ensure that the modeling of the variable is done at the right level of granularity: in other words, to ensure that the effect is associated with the variable itself, and not to a broader class of factors. For instance, Nagel (2013) suggests that a proper corpus study on how the relative frequency of the occurrence of 'know' in children and adults would need to compare it to a "wider range of factive and nonfactive verbs and constructions" (23), presumably to identify which effects belong specifically to 'know', and which ones are spawned by a larger class of verbs.

3 Review

The search for academic work for our review was done opportunistically: some articles were suggested by colleagues, others were found through searches on PhilPapers and Google Scholar, still more were cited by other works that used corpus methods or advocated for them. To be considered as part of this review, a piece had to fulfill the following criteria:

1) It has to be written by philosophers, or at least presented in a philosophical journal or discussed by philosophers. This is because I am interested in the state of corpus methods specifically in philosophy.
2) It has to test a theoretical hypothesis. I therefore excluded descriptive works, and works that formulated non-theoretical hypotheses (e.g. "It is possible to study a concept using corpora")[9].
3) It has to use a representative corpus of textual data to test the hypothesis. This excludes both provoked data[10] (which does not strictly makes for a corpus) and studies where the research hypothesis is about the corpus itself[11].

I interpreted those criteria liberally. For instance, Hinton (2020) states that the main objective of his set of "pilot-studies" is to establish the possibility of using corpora, but it includes the testing of hypotheses, so I chose to include it. On the other hand, I had to overlook works published in languages that I cannot read, like Bluhm's doctoral thesis (2012). In total, I found 27 published works featuring 43 studies, the earliest dating from 2008 with some works still unpublished.

In the rest of this section, I will review this literature and try to assess the practices that have been adopted by practitioners. While I would have liked to assess each work separately, as methods must be adapted to the specific scientific

[9] For this reason, I excluded such works as Alfano, Higgins, and Levernier (2018), Malaterre and Chartier (2019), Malaterre, Chartier, and Pulizzotto (2019), and Dudley (2017).

[10] E.g. Schwitzgebel (2011), Bermúdez et al. (n.d.). Provoked textual data is data obtained by eliciting a response from participants, for example by asking a question or requestning an explanation.

[11] E.g. Sainte-Marie et al. (2011), Danis (2012).

endeavor one is pursuing, this is beyond the scope of this survey. Therefore, after a cursory overview of the selected works, I will try to quantify researchers' use of the tools at their disposal, viz. comparisons and statistical tests. In the last subsection, I will turn to modeling practices. Unlike comparisons and statistical testing, which are easy to identify in an article, most modeling decisions are kept implicit, which complicates the task of documenting them systematically. We can, however, proceed anecdotally and pay a closer look at some of the recurring pitfalls.

3.1 The state of corpus studies in X-Phi

Corpus studies have mainly been employed by experimental philosophers either to explore how scientists and philosophers speak about certain things or to study how certain words or phrases are commonly used. In the first case, they treat the corpus as shedding a light on a community, whereas in the second they use it to study more general features of language and cognition.

Among the first group, we find four papers (Overton 2013; Pease, Aberdein, and Martin 2019; Mejía-Ramos et al. 2019; Mizrahi 2020b) that studied how scientists and mathematicians use terms linked to the concept of explanation. Tallant (2013) and Andow (2015) did similar work with the notion of intuition, and Mizrahi (2020a) worked on the concept of progress. We also find a study by Slingerland and Chudek (2011) on the concept of *xin* in pre-Qin China. Additionally, we find two papers in which the concern is to contrast the way experts use a concept with the way philosophers or laypeople think about it (Napolitano and Reuter, n.d.; Hinton 2020). For this purpose, fairly simple methods are usually sufficient, so these studies usually rely on counting words occurrences and annotating small random samples.

The papers by Murdock, Allen, and DeDeo (2017) and Chartrand and Desmarais Grégoire (2020) are an exception: while they study expert discourse, they are interested in causal factors for the expression of themes and employ more elaborate statistical tools. Schwitzgebel and Jennings (2017) are also an outlier, as they track the uses of pronouns as a proxy for the presence of female authors in philosophical debates.

The studies from the second group, where corpora are a window to human linguistic behavior in general, offer more diversity in themes and methods. We find studies on folk epistemology (Dudley et al. 2017; Nichols and Pinillos 2018; Hansen, Porter, and Francis 2019), evaluative concepts (Wright et al. 2016; Liao, McNally, and Meskin 2016; Reuter, Baumgartner, and Willemson, n.d.), gender and race (Herbelot, Von Redecker, and Müller 2012), perception concepts (Knobe and Prinz 2008; Reuter 2011; Fischer, Engelhardt, and Herbelot 2015; Fischer and Engelhardt 2017) and disposition (Vetter 2014). Unlike studies in the first group, they are often interested in the genesis of concepts, which they often study using data from CHILDES (Wright et al. 2016; Nichols et al. 2016; Dudley et al. 2017; Nichols and Pinillos 2018). They also typically use more varied methodologies, employing machine learning techniques like word vectors (Fischer, Engelhardt, and Herbelot 2015; Sytsma et al. 2019) or clustering algorithms (Reuter, Baumgartner, and Willemson, n.d.) or mixing in traditional experimental methods (e.g. Wright et al. 2016; Nichols et al. 2016; although Napolitano and Reuter, n.d. also present vignette-based experiments).

In both groups, we find a majority of studies that test the association between two variables; but we also find many "non-associative" studies, whose hypotheses involve the characterization of a single variable. More often than not, these non-associative studies make hypotheses about the meaning of words or phrases in their corpus or sample. For instance, Vetter (2014) shows that what folks mean by "to be disposed to" is quite different from the philosophical concept of disposition, and the study in Fischer and Engelhardt (2017) corroborates their hypothesis that the visual meaning for 'see' is much more frequent than the others.

One surprise here is that these studies tend to focus on concepts, or on words that are strongly associated with concepts of interest—this would be unusual among digital humanities studies. Perhaps because of philosophy's history with intuition methods, they also tend to use annotations significantly more (19 out of 43 corpus studies did so).

3.2 Comparisons and statistical testing

In order to get a sense of how likely the authors of our sampled studies were to use the tools at their disposal to strengthen the evidence they produce, I counted how often they used comparisons to control for confounds[12] and how often they used statistical tests. It should be noted here that only mere *use* was documented—not whether it was appropriate or well done.

I found that only 13 studies out of 43 introduced some sort of comparison to control for confounds[13]. In most cases, access to comparable data is probably not the issue. On the one hand, while a few studies in our review used special corpora for which there are few equivalents (like CHILDES data or ancient Chinese literature), most used corpora for which comparable alternatives were easily

[12] As opposed to comparisons used merely to model the variation of an independent variable. For example, say two researchers are interested in whether the concept of love is more about relationship or more about feeling, and both use corpora from China and Brazil. Researcher A is looking at how the country of origin determines how the concept of love is expressed, and thus the multiplicity of corpora is used to model the independent variable, and not to control for potential confounds. Researcher B, however, looks at how the age of the writer influences their concept of love, and they repeat their analysis on both corpora to ensure that their conclusion is valid in both contexts—thus, they use the multiplicity of corpora to control for potential confounds such as culture and language.

[13] To provide a comparison, I looked at the latest issue (volume 17, issue 3, November 2021) from *Corpus Linguistics and Linguistic Theory*, one of the top corpus linguistics journals. Out of the 8 papers in that issue, 7 presented corpus evidence for a linguistic hypothesis—the other is a defense of Structural Equation Modeling for corpus linguistics. Of those 7, 4 made use of comparisons to control confounds (57%) and 6 made use of statistical tests (86%).

available—for example, one might contrast the Corpus of Contemporary American English (COCA) with the British National Corpus (BNC). On the other hand, researchers have the option of observing concepts, words, and linguistic devices of the same category as the ones they are already documenting—and this is completely free: indeed, it is one of the selling points of corpus data[14]. Thus, it would seem that there might be some space for improvement. This said, corpus practitioners have been using more comparisons recently, with most studies since 2019 having done so (figure 1).

Statistical testing is more common, being used in 17 of the 43 studies, including 12 of the 26 associative studies, but given that it is the norm both in experimental philosophy and in corpus linguistics[15], this figure still seems somewhat low. This is not completely unexpected, because, as I mentioned in the previous section, statistical testing on corpora, where researchers control very little, is much harder than it is on experiments. The problem is, however, that where statistical testing was not performed, most researchers did not pursue other means of ensuring that the observed effect was not due to chance. Furthermore, while there has been a slight uptick in the last three years, the use of statistical has remained relatively constant across the last decade (figure 1).

3.3 Modeling variables in textual data

A closer look at the selected articles reveals that there were a few recurrent issues with the modelization of variables. The first one concerns the passage from the research to the experimental hypothesis. We have noted that in many studies, researchers formulated their experimental hypothesis as being about a single variable, rather than the interaction between two variables, and went on to design a non-associative study. While it often seems warranted, there are several cases where the research hypothesis suggests an association between two variables, but the independent variable has vanished from the experimental hypothesis. Consider for example Mizrahi's (2020a, 4) paper, which makes an interesting proposition: we can use corpora to test theories about a concept by looking at the topics that are prompted when someone talks about that concept. He makes the following research hypothesis:

> "On the semantic account, which defines scientific progress in terms of truth, we would expect to find that practicing scientists talk about scientific progress in terms of truth more than knowledge or understanding in scientific publications."

This hypothesis suggests that "talk about progress" is what conditions the appearance of discourse about truth, so we would expect a control condition where

[14] For example, Sytsma et al. (2019) measured the positive/negative/neutral valence of words that were the object of the verb "to cause" in COCA. When they observed that the most frequent words were overwhelmingly negative, they looked at various synonymous verbs and expressions to see if the phenomenon was limited to "to cause" or if it was a more general feature of language.

[15] Cf. note 12 above.

discourse is about something other than progress. His experimental design, however, consists in looking for phrases that express truth, knowledge, or understanding as goals, regardless of the context. Perhaps Mizrahi is assuming that scientists are always, somehow, talking about progress, but if so, there is still a need for a control condition where the topic is not progress, if only to ensure that truth is indeed more common than it usually is and that the occurence frequencies in his corpus aren't, say, simply reflecting how much those words occur in language.

Mizrahi is not the only one to fall into this trap. Several of the other studies restrict their observations to a small set of text segments that contain a specific concept or phrase, and in some cases, it seems that modeling this constraint into a variable would have given more reliable results. This is true even in some of the most compelling works, as in Hansen, Porter, and Francis (2019): after reporting exploratory work on the word 'know', their first quasi-experimental study addresses a claim by Bach (2005), according to which when we deny knowledge to someone, we usually mean to deny that they have a true belief. To test this, they sample instances of denials of knowledge, and find that indeed, most of them are actually denials of true belief. However, this falls short of establishing an association. It could very well be that there are no true beliefs to deny in the first place: Bach himself suggests that "ordinarily, we do not already assume that [knowers] have a true belief and just focus on whether or not their epistemic position suffices for knowing." (2005, 62–53, as cited in Hansen et al. 2019, 256). To have evidence that the negation denies true belief, we need to look at what happens when there is no negation.

Given that many of the reviewed studies deal with concepts, another issue is with the modeling of concepts in textual data. Words are not concepts, and in natural speech and writing words are usually highly polysemic, while concepts are expressed through a variety of expressions. The onus is therefore on the researcher to model their hypotheses so as to ensure that their corpus can corroborate them.

As illustrated across the studies on the concept of explanation, the challenge posed by the word-concept gap varies a lot depending on the research question being asked. The first study by Overton (2013, 1386) looks at the importance of this concept in scientific discourse. Recognizing the challenge in translating hypotheses about concepts to the lexical level, he proposes this principle: "If explain is a more important concept than others such as cause, then we would expect "explain" words to occur more frequently than "cause" words." Once we accept it, formulating the experimental hypothesis is trivial. The second study by Pease, Aberdein, and Martin (2019) explores different aspects of the concept of explanation as it is used, so Overton's stratagem is not available to them. Therefore, they code occurrences of explanation words to ensure that they referred to the right concept. The question Mejía-Ramos et al. (2019) are interested in is whether and how mathematicians describe themselves as explaining, which also sidesteps some of the problems of modeling concepts with words. However, they seem to implicitly assume that using explanation words and describing oneself as explaining are the same—one might think that this assumption should be motivated. Finally, the approach adopted by Mizrahi (2020b) is the one that raises the most concerns, as it interprets the use of words like "explain" not even as a reference to the concept of explanation, but as indicating that the text excerpt that contains this segment is an explanation. But obviously people do not label their discourse as they are engaging in it. Mizrahi comes close to providing a compelling answer to the word-concept gap when he notes that argumentation textbooks often list markers of arguments: this approach could be promising, given that keyword approaches have been quite successful in argument mining (Lawrence and Reed 2015). However, he rejects them for being too indiscriminating—a test to which his own approach is not confronted.

Furthermore, one would expect absent or faulty modeling to be at the root of other problems. For instance, modeling has implications down the line for comparisons and statistical testing[16]. Good models will make it much easier to identify confounds, and thus to choose the right comparisons to produce strong evidence. Conversely, faulty modeling will make it harder to identify the right null hypothesis, and thus also make it harder to find the right statistical test. Therefore, one might expect that better and more thorough modeling would translate into more and better comparisons and statistical tests.

4 Discussion

Corpus methods have advertised themselves in philosophy by the potential that they could unleash: they open up new ecologically valid observables against which theories can be tested (Ulatowski et al. 2020), they provide tools for the discovery of new hypotheses (Caton 2020), an alternative to intuition (Bluhm 2016) and an access to the minds of the past (Slingerland and Chudek 2011) and of other cultures.

Fulfilling this potential, or at least some of it, is necessary if corpus methods are to take their place in experimental philosophy. Because they are marginal, and because corpora are very complex datasets, most philosophers lack the knowledge to master and assess corpus methods. As a result, corpus methods provoke reactions that range from the overly optimistic enthusiasm of researchers who see the possibilities, but not the obstacles, to the excessive caution or pessimism from colleagues who feel that they shouldn't trust methods that have yet to demonstrate their reliability. Yet, both the enthusiasm and the caution are warranted. It is therefore up to corpus practitioners to meet the methodological challenges and give philosophers reasons to trust corpus studies.

In that regard, current practices could be improved. Practitioners often fail to make the comparisons that could control for confounds, and omit statistical

[16] I'm indebted to Jean-Guy Meunier for this observation.

testing where it would be relevant. We have also seen that, although we cannot pronounce ourselves on the prevalence of this problem, there are cases where modeling practices are inadequate. However, this is not necessarily an indictment of quasi-experimental corpus studies in philosophy. On the one hand, a less compelling methodological choice does not make for a bad study—indeed, the studies scrutinized in this paper provide novel and interesting takes on their objects of study, often drafting the contours of their textual footprint for the very first time. Shedding a light on shortcomings certainly takes nothing away with regards to the argumentation, impact, or originality of these works.

On the other hand, one might argue that corpus methods are still new on the experimental philosophy scene and that as practitioners become more comfortable with the relevant methods[17], we might see better practices being integrated. As a matter of fact, the recent uptick in the use of comparisons and statistical tests is a good omen. However, this progress would likely be helped by a discussion of the methodological issues. This paper is meant as a step in this direction.

Furthermore, from our review, we learn that some practices could probably do much to improve philosophers' corpus methods[18]:

1. Practitioners should seek to justify their intuitions about how their object will manifest itself in textual data. For instance, if one is to study how mathematicians use arguments, they should ensure that their text markers are adequate to detect those arguments. If necessary, and if human annotation is unavailable or scarce, they could also turn to machine learning techniques, such as topic models (as Chartrand and Desmarais Grégoire 2020 have done with discussions of testimony) or classifiers (e.g. Lawrence and Reed 2015).

2. Practitioners should look to include comparisons as much as is possible. In particular, they should ensure that they look at baseline or controls conditions rather than only at the condition they are interested in. For instance, philosophers interested in knowledge denials should also look at other types of pragmatic contexts. They should also try to include more than one corpus.

3. Practitioners should seek experimental designs that make it possible to

[17] Given that their objects differ from those studied by linguists or digital humanists, it is likely that philosophers will develop corpus methods specifically suited to their aims. In the meantime, it would probably be a good idea to look at corpus linguistics for inspiration (e.g. Paquot and Gries 2020, pt. IV), as they tend to pursue project that are methodologically similar to those of experimental philosophers. However, cf. Baroni and Evert (2009) for a great introduction on relevant statistical methods by computational linguists.

[18] There are, however, other practices that could be very valuable. An anonymous reviewer suggested preregistration: while it raises special challenges in the context of NLP and corpus studies, it is a very promising avenue (cf. Miltenburg, Lee, and Krahmer 2021). Another approach could be to create baselines with randomized data—a practice that is well-established in computational linguistics, and which encourages scientists to develop a model of their assumptions concerning the generation of textual data and of how they believe their hypothesis fits within this picture.

make statistical tests. For instance, by introducing variation in the independent variable, Hansen, Porter, and Francis (2019) could have turned their experiment into a measure of the association between knowledge denials and the uses of the word ‘know’, and could thus have tested for the significance of the association between knowledge denials and absence of belief.

5 Conclusion

In this article, I reviewed existing works in corpus quasi-experiments in philosophy, with a particular focus on an issue that seems to be a sticking point in this literature: modeling and design practices. What emerged was that comparisons and statistical tests are relatively rare in this body of work (much more than in corpus linguistics, for example), and that, at least in some cases, researchers make modeling decisions that either do not take into account the nature of corpora and of the word-concept relationship or undermine the experiment’s capacity to answer research questions. Thus, we provided a glance at this body of literature, and emphasized pitfalls that future researchers can strive to avoid.

The upshot of this is that there is much to be gained by philosophers in developing and adopting best practices in testing philosophical hypotheses using text corpora. As we have shown here, proficiency with corpus methods is not only proficiency with the computational tools involved in it, but also an understanding of the way they are used to produce evidence.

6 Acknowledgements

I wish to thank Édouard Machery, Kevin Reuter, Justin Sytsma, and an anonymous reviewer for their invaluable comments and suggestions, as well as Jean-Guy Meunier for his insights. Louis Chartrand is supported in part by funding from the Social Sciences and Humanities Research Council.

7 References

Alfano, Mark. 2018. “Digital humanities for history of philosophy: A case study on Nietzsche.” In *Research Methods for the Digital Humanities*, 85–101. Springer.

Alfano, Mark, Andrew Higgins, and Jacob Levernier. 2018. “Identifying Virtues and Values Through Obituary Data-Mining.” *Journal of Value Inquiry* 52 (1). https://doi.org/{10.1007/s10790-017-9602-0}.

Andow, James. 2015. “How “Intuition” Exploded.” *Metaphilosophy* 46 (2): 189–212. https://doi.org/{10.1111/meta.12127}.

———. 2016. "Qualitative Tools and Experimental Philosophy." *Philosophical Psychology* 29 (8): 1128–41. https://doi.org/{10.1080/09515089.2016.1224826}.

Bach, Kent. 2005. "The Emperor's New 'Knows'." In *Contextualism in Philosophy: Knowledge, Meaning, and Truth*, edited by Gerhard Preyer and Georg Peter, 51–89. Oxford University Press.

Baroni, Marco, and Stefan Evert. 2009. "Statistical methods for corpus exploitation." *Corpus linguistics: An international handbook* 2: 777–803.

Bermúdez, Juan Pablo, Samuel Murray, Louis Chartrand, and Sergio Barbosa. n.d. "What's inside is all that matters? The contours of everyday thinking about self-control." *Review of Philosophy and Psychology.*

Bluhm, Roland. 2012. *Selbsttäuscherische Hoffnung: Eine Sprachanalytische Annäherung.* mentis.

———. 2013. "Don't Ask, Look! Linguistic Corpora as a Tool for Conceptual Analysis." In *Was dürfen wir glauben?: Was sollen wir tun? Sektionsbeiträge des achten internationalen Kongresses der Gesellschaft für Analytische Philosophie e. V.*, edited by Migue Hoeltje, Thomas Spitzley, and Wolfgang Spohn, 7–15. DuEPublico.

———. 2016. "Corpus Analysis in Philosophy." In *Evidence, Experiment and Argument in Linguistics and the Philosophy of Language*, edited by Martin Hinton, 91–109. Peter Lang.

Caton, Jacob N. 2020. "Using Linguistic Corpora as a Philosophical Tool." *Metaphilosophy* 51 (1): 51–70. https://doi.org/{10.1111/meta.12405}.

Chartier, Jean-François, Jean-Guy Meunier, Jean Danis, and Mohamed Jendoubi. 2008. "Le travail conceptuel collectif: une analyse assistée par ordinateur du concept d'ACCOMMODEMENT RAISONNABLE dans les journaux québécois." *Heiden, S. et Pincemin, B., editors, JADT*, 297–307.

Chartrand, Louis. 2017. "La philosophie entre intuition et empirie: Comment les études du texte peuvent contribuer à renouveler la réflexion philosophique." *Artichaud Magazine* 2017 (8 juin).

Chartrand, Louis, and Desmarais GrégoireUlisce. 2020. "Testimony in sexual and nonsexual assault trials: conviction is possible." Digital Humanities 2020. {https://hcommons.org/app/uploads/sites/1000360/2020/07/article.pdf}.

Chow, Siu L. 1998. "Précis of statistical significance: Rationale, validity, and utility." *Behavioral and brain sciences* 21 (2): 169–94.

Cova, Florian. 2016. "The Folk Concept of Intentional Action." In, 117–41. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118661666.ch8.

Danis, Jean. 2012. "L'analyse conceptuelle de textes assistée par ordinateur (LACTAO): une expérimentation appliquée au concept d'évolution dans l'oeuvre d'Henri Bergson." Master's thesis, Université du Québec à Montréal.

Dudley, Rachel. 2017. "The role of input in discovering presuppositions triggers: Figuring out what everybody already knew." PhD thesis.

Dudley, Rachel, Meredith Rowe, Valentine Hacquard, and Jeffrey Lidz. 2017. "Discovering the factivity of" know"." In *Semantics and Linguistic Theory*, 27:600–619.

Fischer, Eugen, and Mark Curtis. 2019. *Methodological Advances in Experimental Philosophy.* London: Bloomsbury Press.

Fischer, Eugen, and Paul E. Engelhardt. 2017. "Diagnostic Experimental Philosophy." *Teorema: International Journal of Philosophy* 36 (3): 117–37.

Fischer, Eugen, Paul E. Engelhardt, and Aurelie Herbelot. 2015. "Intuitions and Illusions: From Explanation and Experiment to Assessment." In *Experimental Philosophy, Rationalism, and Naturalism. Rethinking Philosophical Method*, edited by Eugen Fischer and John Collins, 259–92. Routledge.

Forest, Dominic, and Jean-Guy Meunier. 2000. "La classification mathématique des textes: un outil d'assistance à la lecture et à l'analyse de textes philosophiques." In *Proceedings of JADT*, 325–29.

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

Gries, Stefan Th. 2005. "Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff." *Corpus Linguistics and Linguistic Theory* 1 (2): 277–94. https://doi.org/10.1515/cllt.2005.1.2.277.

Hansen, Nat, J. D. Porter, and Kathryn Francis. 2019. "A Corpus Study of "Know": On the Verification of Philosophers' Frequency Claims About Language." *Episteme*, 1–27. https://doi.org/10.1017/epi.2019.15.

Herbelot, Aurélie, Von RedeckerEva, and Johanna Müller. 2012. "Distributional techniques for philosophical enquiry." In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 45–54.

Hinton, Martin. 2020. "Corpus Linguistics Methods in the Study of (Meta)Argumentation." *Argumentation*, 1–21. https://doi.org/10.1007/s10503-020-09533-z.

Hitchcock, Christopher. 2020. "Causal Models." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2020. Metaphysics Research Lab, Stanford University. {https://plato.stanford.edu/archives/sum2020/entries/causal-models/}.

Kilgarriff, Adam. 2005. "Language is never, ever, ever, random." *Corpus Linguistics and Linguistic Theory* 12: 263–75.

Knobe, Joshua. 2003. “Intentional action and side effects in ordinary language.” *Analysis* 63 (3): 190–94.

Knobe, Joshua, and Jesse Prinz. 2008. “Intuitions About Consciousness: Experimental Studies.” *Phenomenology and the Cognitive Sciences* 7 (1): 67–83. https://doi.org/{10.1007/s11097-007-9066-y}.

Koplenig, Alexander. 2019. “Against statistical significance testing in corpus linguistics.” *Corpus Linguistics and Linguistic Theory* 15 (2): 321–46.

Lakatos, Imre. 1970. “Falsification and the Methodology of Scientific Research Programmes.” In *Criticism and the Growth of Knowledge*, edited by Imre Lakatos and Alan Musgrave, 91–195. Cambridge University Press.

Lawrence, John, and Chris Reed. 2015. “Combining argument mining techniques.” In *Proceedings of the 2nd Workshop on Argumentation Mining*, 127–36.

Liao, Shen-yi, Louise McNally, and Aaron Meskin. 2016. “Aesthetic Adjectives Lack Uniform Behavior.” *Inquiry: An Interdisciplinary Journal of Philosophy* 59 (6): 618–31. https://doi.org/{10.1080/0020174x.2016.1208927}.

Malaterre, Christophe, and Jean-François Chartier. 2019. “Beyond Categorical Definitions of Life: A Data-Driven Approach to Assessing Lifeness.” *Synthese*. https://doi.org/{10.1007/s11229-019-02356-w}.

Malaterre, Christophe, Jean-François Chartier, and Davide Pulizzotto. 2019. “What is This Thing Called Philosophy of Science?: A Computational Topic-Modeling Perspective, 1934–2015.” *Hopos: The Journal of the International Society for the History of Philosophy of Science* 9 (2): 215–49. https://doi.org/10.1086/704372.

McKinnon, Alastair. 1970. *The Kierkegaard Indices.* Leiden: Brill.

Mejía-Ramos, Juan Pablo, Lara Alcock, Kristen Lew, Paolo Rago, Chris Sangwin, and Matthew Inglis. 2019. “Using Corpus Linguistics to Investigate Mathematical Explanation.” In *Methodological Advances in Experimental Philosophy*, edited by Eugen Fischer and Mark Curtis, 239–63. London: Bloomsbury Academic.

Meunier, Jean-Guy. 2017. “Theories and Models: Realism and Objectivity in Cognitive Science.” In *Varieties of Scientific Realism*, 331–52. Springer.

Meunier, Jean Guy, Dominic Forest, and Ismail Biskri. 2005. “Classification and Categorization in computer-assisted reading and text analysis.” In *Handbook of categorization in cognitive science*, 955–78. Elsevier.

Meunier, Jean-Guy, Stanislas Rolland, and François Daoust. 1976. “A system for text and content analysis.” *Computers and the Humanities*, 281–86.

Miltenburg, Emiel van, Chris van der Lee, and Emiel Krahmer. 2021. “Preregistering NLP research.” In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, 613–23. Online: Association for Computational Linguistics. https://doi.org/{10.18653/v1/2021.naacl-main.51}.

Mizrahi, Moti. 2020a. “Conceptions of scientific progress in scientific practice: an empirical study.” *Synthese*, 1–20.

———. 2020b. “Proof, explanation, and justification in mathematical practice.” *Journal for General Philosophy of Science*, 1–18.

Murdock, Jaimie, Colin Allen, and Simon DeDeo. 2017. “Exploration and Exploitation of Victorian Science in Darwin’s Reading Notebooks.” *Cognition* 159: 117–26. https://doi.org/{10.1016/j.cognition.2016.11.012}.

Nagel, Jennifer. 2013. “Knowledge as a Mental State.” *Oxford Studies in Epistemology* 4: 275–310.

Napolitano, M. Giulia, and Kevin Reuter. n.d. “What is a Conspiracy Theory?”

Nichols, Shaun, Shikhar Kumar, Theresa Lopez, Alisabeth Ayars, and Hoi-Yee Chan. 2016. “Rational Learners and Moral Rules.” *Mind and Language* 31 (5): 530–54. https://doi.org/{10.1111/mila.12119}.

Nichols, Shaun, and N. Ángel Pinillos. 2018. “Skepticism and the Acquisition of “Knowledge”.” *Mind and Language* 33 (4): 397–414. https://doi.org/{10.1111/mila.12179}.

Overton, James A. 2013. ““Explain” in Scientific Discourse.” *Synthese* 190 (8): 1383–1405. https://doi.org/{10.1007/s11229-012-0109-8}.

Paquot, Magali, and Stefan Thomas Gries. 2020. *A practical handbook of corpus linguistics*. Springer Nature.

Pease, Alison, Andrew Aberdein, and Ursula Martin. 2019. “Explanation in Mathematical Conversations: An Empirical Investigation.” *Philosophical Transactions of the Royal Society A* 377.

Pettit, Dean, and Joshua Knobe. 2009. “The pervasive impact of moral judgment.” *Mind & language* 24 (5): 586–604.

Reuter, Kevin. 2011. “Distinguishing the Appearance From the Reality of Pain.” *Journal of Consciousness Studies* 18 (9-10): 94–109.

Reuter, Kevin, Lucien Baumgartner, and Pascale Willemson. n.d. “Tracing Thick Concepts Through Corpora.”

Sainte-Marie, Maxime B., Jean-Guy Meunier, Nicolas Payette, and Jean-François Chartier. 2011. “The concept of evolution in the Origin of Species: a computer-assisted analysis.” *Literary and linguistic computing* 26 (3): 329–34.

Schwitzgebel, Eric. 2011. *Perplexities of Consciousness*. Bradford.

Schwitzgebel, Eric, and Carolyn Dicey Jennings. 2017. “Women in Philosophy: Quantitative Analyses of Specialization, Prevalence, Visibility, and Generational Change.” *Public Affairs Quarterly* 31: 83–105.

Slingerland, Edward, and Maciej Chudek. 2011. “The Prevalence of Mind–Body Dualism in Early China.” *Cognitive Science* 35 (5): 997–1007. https://doi.org/{10.1111/j.1551-6709.2011.01186.x}.

Sytsma, Justin, Roland Bluhm, Pascale Willemsen, and Kevin Reuter. 2019. “Causal Attributions and Corpus Analysis.” In *Methodological Advances in Experimental Philosophy*, edited by Eugen Fischer. Bloomsbury Press.

Tallant, Jonathan. 2013. “Intuitions in Physics.” *Synthese* 190 (15): 2959–80. https://doi.org/{10.1007/s11229-012-0113-z}.

Ulatowski, Joe, Dan Weijers, Justin Sytsma, and Colin Allen. 2020. “Cognitive Science of Philosophy Symposium: Corpus Analysis.” Edited by Zina Ward. *The Brains Blog*. The Brains Blog. {https://philosophyofbrains.com/2020/12/15/cognitive-science-of-philosophy-symposium-corpus-analysis.aspx}.

Vetter, Barbara. 2014. “Dispositions without conditionals.” *Mind* 123 (489): 129–56.

Wright, Jennifer Cole, Trisha Sedlock, Jenny West, Kelly Saulpaugh, and Michelle Hopkins. 2016. “Located in the Thin of It: Young Children’s Use of Thin Moral Concepts.” *Journal of Moral Education* 45 (3): 308–23. https://doi.org/{10.1080/03057240.2016.1156523}.

8 Tables and figures

Table 1: Breakdown of hypotheses for Knobe (2003) following the framework by Chow (1998), with E being a decription of the experimental apparatus.

Level of discouse	Assumes	Assertion	Identifier
To be explained		The concept of intentional action	
Substantive theory		The concept of intentional action is sensitive to valence when it comes to side-effects	$P1$
Complement of substantive theory		The concept of intentional action is insensitive to valence when it comes to side-effects	$P1'$
Research hypothesis	$P1$	Ascriptions of intentionality for side-effects are sensitive to valence	$P2$
Complement of research hypothesis	$\neg P1$	Ascriptions of intentionality for side-effects are insensitive to valence	$P2'$
Experimental hypothesis	$P2$	People are more likely to say a side-effect was produced intentionally if it is harmful than if it is helpful	$P3$
Complement of experimental hypothesis	$\neg P2$	People are as likely to say a side-effect was produced intentionally if it is harmful as if it is helpful	$P3'$
Statistical alternative hypothesis	$P3$, E	Rate of intention ascriptions in the harm condition $>$ rate of intention ascription in the help condition	h_1
Statistical null hypothesis	$\neg P3$, E	Rate of intention ascriptions in the harm condition $\leq$ rate of intention ascription in the help condition	h_0

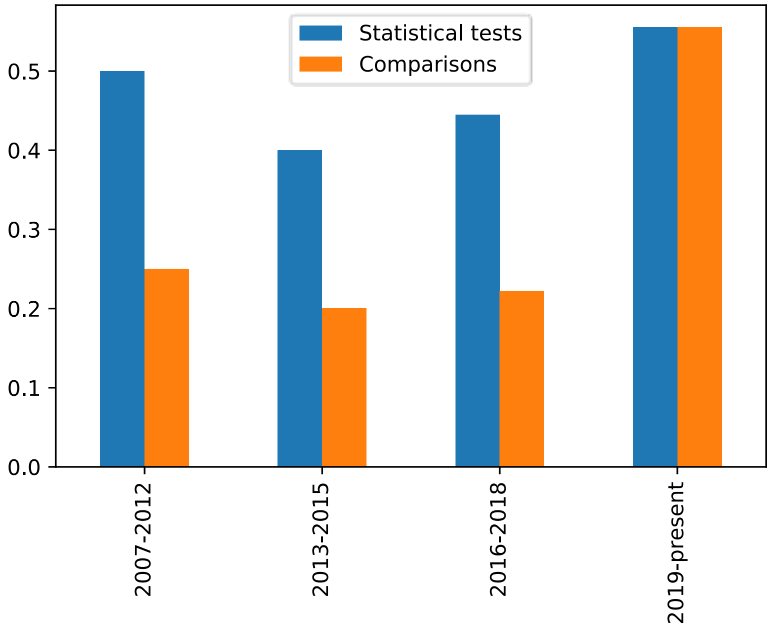

Figure 1: Use of statistical tests and confound-controling comparisons, 2011-present.