

Roland Bluhm

# Corpus Analysis in Philosophy

## 1. Introduction

The aim of this paper is to discuss the potential benefit of corpus analysis, a (partly) empirical method from linguistics, for philosophy. ‘Corpus analysis’ is not only the name of the method, but also a rough description of it, because the method consists in analysing data taken from linguistic text corpora. In linguistics, using such text corpora is an established practice. A fair number of them are nowadays freely accessible on the internet, and using them has become relatively easy, even for researchers without linguistic expertise. Surprisingly, corpus analysis has been widely disregarded by philosophers, including those that profess a methodical interest in language—a state of affairs that I believe ought to change.

I shall begin with a short introduction into corpus analysis, followed by a couple of general remarks on why we should, in my view, use corpus analysis in philosophy. I shall briefly introduce examples of the little work that has been done with the help of corpus analysis in philosophy, which I shall contrast with the use of general internet search engines and questionnaires for similar purposes. I shall close with some remarks on the advantages and disadvantages of corpus analysis in philosophy, and some suggestions for directions that further philosophical research with the help of corpus analysis might take.

## 2. Basics of corpus analysis

In order to introduce corpus analysis, it would be helpful to say what a linguistic text corpus is; but, regrettably, I have yet to find a completely satisfactory definition. Bluntly phrased, a text corpus is just a heap of texts; but, and here the problems lurk, not any old heap of texts is supposed to be a linguistic text corpus. For the purpose of this paper, I would like to evade the nicer difficulties by relying on the following working definition: a linguistic text corpus is, roughly speaking, (i) a collection of (written or spoken) texts that (ii) serves as the primary data base for answering language-related research questions, and (iii) has been collected and structured for this purpose, or is at the very least considered for this purpose.

This characterisation suffices for my purposes, although it is not wholly satisfactory. Let me hint at some of its problems. Difficulties pertain to the third criterion, because it is doubtful whether a precise line can be drawn between principled and

unprincipled collections of texts.<sup>1</sup> This also raises the issue of balance: a linguistic text corpus should contain a suitable choice of texts, and suitability is not only relative to the use to which the corpus is put, but also to questions of representativeness. A useful corpus should be computerised and annotated, that is, in addition to the content of the texts it should contain information, e.g. on the source of the texts, but also grammatical information on words.<sup>2</sup> And in order to access corpus data, a suitable software tool is required. Clearly some of these properties are normative demands, characterising usefulness instead of corpusness, but I shall shirk further debate and instead rely on my working definition.

There are many linguistic text corpora available. I do not have the space to introduce any number of them, let alone in detail.<sup>3</sup> Two corpora that are frequently mentioned beyond the boundaries of linguistics are the British National Corpus (abbreviated “BNC”) and the Corpus of Contemporary American English (“COCA”). The BNC is a relatively large, closed corpus of texts of written and spoken language. It contains approximately 100 million words in texts dating from 1960 to 1994. The Corpus of Contemporary American English (COCA) is not closed; every year approximately 20 million words are added. At present, the corpus comprises more than 520 million words, in about 220,000 texts dated from 1990 to the present, the most recent stemming from December 2015. COCA has a quite well-balanced collecting policy, and it is 20% spoken language. Both BNC and COCA are freely accessible for scientific purposes, and there is a comfortable web interface for their use provided by the Brigham Young University in Provo, Utah.<sup>4</sup>

The basic idea of using corpora is to have controlled access to empirical linguistic data—empirical in the sense that it is language that has in fact been used. Having controlled access to the data of a computerised corpus on the most fundamental level means that a search engine can be used to query for an expression in the texts collected in the corpus. The search is executed very quickly, and the occurrences of the expression in the corpus are numbered and listed in a so-called KWIC index (Key Word In Context, *i.e.*, the queried expression with short extracts

---

1 Cf. Hundt (2008: 170).

2 Although corpora do not need to be computerised, it has to be noted that the development of corpus linguistics is not only intimately tied to the development of computer technology, but that the increase in the quantity of data that can be processed has brought about qualitative differences in our methodical accessing of the data as well (cf. Bonelli 2010: 15–18).

3 Cf. Lee (2010) or Xiao (2008) for an overview.

4 At present: <<http://corpus.byu.edu/>>. The interface gives access to a number of other corpora, too.

of the context preceding and following it). Usually, more of the context in which the key word was found can be accessed. The available search algorithms are, of course, much more powerful, and I am going to say more about them later. What is important to me here only is to point out the most basic principle of corpus analysis.

It should be obvious that handling data extracted from a corpus can be a lot of work. For example, searching COCA for the expression 'hope' yields 92,467 hits. In order to extract anything useful out of this amount of data, it needs to be processed somehow, and the sheer number of occurrences hints at the effort this may require. Given that this is so, the natural question to ask is: why bother? Why use corpora in philosophy at all?

### 3. Why use corpora in philosophy?

Quite obviously, using corpora for philosophical purposes makes sense only if it matters how things are represented in language. This need not necessarily mean a natural language or the ordinary use or variety of a natural language, but in what follows I shall be mostly concerned with natural language in its ordinary use.<sup>5</sup> Let me also state explicitly that I shall not offer any general justification for considering ordinary language in philosophy, but am only concerned with giving reasons for using corpora for this purpose.

Suppose that you have a research interest and some (perhaps tacit) hypotheses related to the issue that you want to address. If linguistic phenomena are important, it is still usually unclear exactly which phenomena are to be considered. For example, often the aim of linguistic analyses in philosophy is to analyse concepts; but concepts can usually be expressed in a given language in various ways. It is, therefore, not automatically clear, which linguistic phenomena are pertinent for the analytical process. Thus, the first step of linguistic analyses in philosophy is to clarify which expressions are to be considered at all. Only then can one form (again, perhaps tacit) hypotheses about their use. And only then can one test and refine these hypotheses by coming up with examples (linguistic contexts or hypothetical cases), by interpreting them, and examining or expanding on potentially interesting findings through iterations and variations. Ideally, this process ends with conclusions being drawn from testing and refining, with the initial hypotheses falsified, refined, or corroborated.

---

5 A corpus of utterances in an artificial language is odd only if we assume that there is no pragmatic context and no possibility to adapt the language to purposes that go beyond its rules. If the language is *used*, a corpus may provide insights.

This simplified description is sufficient to bring out that philosophical analyses of natural language involve intuitions in two senses of the word. On the one hand, intuitions in the sense of spontaneous non-inferential judgements, marked by a high degree of subjective certainty, that concern the acceptability of given uses of linguistic expressions (in linguistic or pragmatic contexts). On the other hand, intuition in the sense of the faculty or ability on which judgements of the said kind rest, *viz.* competence in the object language. A linguistic analysis involves intuition in both senses. It requires (at various stages of the process) judgements about the acceptability of the use of an expression in a given context, and it requires linguistic competence. In a *passive* sense, *i.e.*, in the sense of the mere understanding of the language, this competence is required at all stages of an analysis, since you have to be able to *understand* what you are analysing (which is the reason why the method is only partly empirical), but in order to think of pertinent linguistic expressions and to construe examples of their use, an *active* competence in the object language is required. Thus, there is an active and a passive role of (the faculty of) intuition in philosophical analyses of linguistic phenomena.

Relying on intuition, particularly in the active sense, is problematic in two ways. First, the faculty of intuition is relative to the individual, and everyone's linguistic competence is limited, especially the active part of it: just think of the difference between your passive and your active vocabulary. For this reason, intuition is only a limited source of data.<sup>6</sup> It is worth mentioning at this point that philosophers who are not native speakers of English have an even more limited linguistic competence. This is relevant because English at the moment is more often than not the object language of philosophical research. A second problem in relying on intuition is the danger of bias. The one who thinks of the examples usually has an investment in specific hypotheses, which might bias the choice and the interpretation of the examples. Of course, such human weakness rarely does befall philosophers, but even we occasionally do fall into the mental traps of mere mortals.

One might object that my picture of what is usually done in linguistic analysis in philosophy is too rough. I readily admit to considerable simplification. Most important for our purposes, there are a number of ways in which philosophers have tried to overcome the limits of their own intuition. This is what I like to call arming the armchair, and there are two time-honoured ways of doing so. The first

---

6 Haß (1991: 230) points out that for this reason intuition cannot serve as a primary database. He also emphasizes that competence in the sense of the linguist's trained understanding of language is necessary to derive hypotheses from empirical data.

is to ask a colleague or anyone else with respectable linguistic competence and sufficient patience what she or he thinks—that is, to ask an (expert) informant. The second is to peruse a dictionary. The aims of using a dictionary are quite similar to those of asking informants. One is to identify the pertinent expressions for an analytical task at hand, another is to get an account of the presumed meaning(s) of these expressions. Yet another is to be given examples for their use in ordinary language.

However, neither dictionaries nor expert informants are to be trusted unquestioningly. First of all, neither dictionaries nor informants are without error. Some mistakes may be idiosyncratic; others may have systematic causes. For example, systematic mistakes of informants may be due to the limits of their knowledge of the object language. And with respect to dictionaries, we have to keep in mind that all dictionaries need to choose some material over other material. Not everything *can* be recorded; and what *is* to be recorded is a question of the policy of the dictionary. This is related to a second problem: both informants *and* dictionaries rely on intuition at various points, the intuition of authors and editors, but, especially in older dictionaries, also the intuition of informants. Thirdly, new dictionaries partly rely on older dictionaries. Dictionary writers copy to some degree what other dictionary writers have written. This is hardly avoidable for reasons of economy, and it is also a reasonable thing to do. It is an academic virtue to preserve knowledge that has already been gained, but, regrettably, when we consult a dictionary, we do not know which material the editors of the dictionary have simply inherited from their predecessors. Fourthly, although dictionaries may be helpful for formulating preliminary hypotheses about meanings, the definitions sought in philosophy differ in function, focus and degree of precision from the paraphrases of meanings given in lexicography.<sup>7</sup>

If we want to avoid these potential shortcomings of informants and of dictionaries, we need some basis on which we can test, correct and extend their claims. More particularly, we need independent, and thus unbiased, evidence that the expressions we have an interest in are used in certain ways. And we need this independent evidence to test our hypotheses about the use of these expressions.

The solution is, I believe, to turn to linguistic text corpora, because they fulfil these desiderata—at least if chosen correctly. We can assume that corpus data that has been recorded and compiled independently is unbiased with respect to specific research questions.<sup>8</sup> But the *usefulness* of a corpus for a given philosophical purpose

---

7 Cf. Wiegand (1989).

8 Cf. Schütze (2010: 117).

has further prerequisites. I do not want to go into much detail, but it should come as no surprise that the normative criteria for corpora resurface here. The first prerequisite is that the data contained in the corpus has to be *suitable* for the specific purpose for which it is to be used. For example, it does not make much sense to use a corpus of academic books to study philosophical terms as they are used in ordinary language. Generally speaking, the corpus has to be *balanced*, that is, the database must contain carefully weighted text types, so as not to give skewed evidence of language use. It has to have *sufficient size* to contain (enough) relevant data. Size is not a value in itself; for some purposes a small corpus will suffice (e.g. the study of a specific author's work), but the more infrequent the phenomenon of interest, the bigger the corpus has to be. If nuanced and extensive interpretation is required, the corpus also has to contain a significant part of the *context* of the queried expressions. Some corpus tools clip these contexts due to copyright restrictions. While searches can be executed over the complete data, only extracts of the data are accessible for analysis. Another prerequisite of the usefulness of a corpus is that it is *computerised* and *annotated*, that is, the word tokens must have been analysed grammatically and semantically and have been tagged accordingly, e.g. with their grammatical categories. Only then can the corpus support sophisticated search algorithms and thus support powerful *software tools* to access the corpus data.

I would like to expand on the issue of annotation and search algorithms, but before I do so, let me mention that BNC and COCA meet these prerequisites, as do many other (but by no means all) corpora.

#### 4. How can corpora be used in philosophy?

We have moved on to the somewhat more practical questions: *how* can corpora be used in philosophy, and *when* should they be used? There is no exhaustive answer to these questions, but a picture of the possibilities can be extrapolated from the search options that standard corpus software offers in conjunction with examples of the application of corpus analysis in philosophy.

There is a variety of software tools for the use of corpora and there are different search interfaces for corpora accessible through the internet.<sup>9</sup> Some of the available search options are obvious and common, some go well beyond what the non-expert might expect. The most important linguistic phenomena that can be queried for with suitable corpus software or corpus search engines, are: (1) Words,

---

9 A short overview of four generations of corpus tools is given in McEnery and Hardie (2012: 37–48), for a short account of their differences and their importance see Anthony (2013).

*i.e.*, expressions exactly as typed in, e.g. ‘hope’ or maybe ‘hpoe’, if you are very thorough and would like to check for misspelled occurrences of a word you are looking for. (2) Expressions with wildcards, e.g. ‘hoping\*’, which not only returns occurrences of ‘hoping’, but also the occurrences of all other expressions beginning with ‘hoping’, e.g. ‘hopingly’. (3) Lemmata, *i.e.*, all occurrences of a word, disregarding inflexions. Thus, searching for the lemma ‘hope’ returns not only occurrences of the expression ‘hope’, but also of the inflected forms ‘hopes’, ‘hoped’, and ‘hoping’. (4) Expressions of named grammatical categories, e.g., occurrences of ‘hope’ as a noun *vs.* occurrences of ‘hope’ as a verb. (5) Co-occurrence of expressions (within a specified distance of some number of words, and sometimes within a sentence, as defined by punctuation, or within a specified distance of some number of sentences), e.g., you can look for co-occurrence of ‘hopes’ and ‘coffee’ within a sentence in order to find out how often coffee is presented as the object of someone’s hope.<sup>10</sup> (6) Complex expressions, e.g. the lemma ‘hope’ preceded by an adjective. Queries of this type, as well as for type (3), are only possible in annotated corpora.

The uses to which these search options can be put are manifold. I would like to give a handful of examples that will serve to introduce both basic and more sophisticated possibilities.

The first is from my own research on hope. I have tried to answer the descriptive question of what hope is—construed as asking for an explication of the concept in its ordinary language use. And I have tried to answer the normative question, if and how hope can be problematic. In an early phase of my research, I used corpora for explorative purposes, trying to identify ways in which a hope or a person entertaining a hope is criticised. For this purpose, I searched for occurrences of the German lemma ‘hoffen’ (the German verb for ‘hope’) in collocation with any adverb and ‘Hoffnung’ (the German noun for ‘hope’) in collocation with any adjective. From the search results I excluded irrelevant finds (that apparently had no normative component) and in this way came up with pertinent candidates. After repeating the process in other corpora, I analysed the occurrences of the word combinations in context and derived a shortlist of expressions that have in fact been used to portrait a hope or a hoper as somehow deficient in ordinary language. This “shortlist” comprises 56 expressions, including, to name just a few in translation, ‘vain’, ‘unrealistic’, ‘exaggerated’, ‘naïve’, ‘childish’, ‘criminal’, ‘foolish’,

---

10 The results of such a query would have to be checked for relevance, because co-occurrences in a sentence are not necessarily tokens of ‘coffee’ being the syntactical object of ‘hope’, as, for example, in the following sentence: “The airline’s coffee is beyond all hope.” (COCA #1)

and ‘immoral’ hope.<sup>11</sup> Even without going into detail, it is possible to appreciate that there are many ways in which hope can be deficient, and that it is possible to track them by looking at the language of hope.

One of my descriptive hypotheses regarding hope was that there are opposite affective experiences of hoping. Traditionally, if hope has been construed as some sort of emotion, it has always been construed as a positive one. To me that seemed to be an inadequate simplification. In order to test my hypothesis, I searched for the lemma ‘hope’ (this time in English) qualified by an adjective or adverb. The results showed that on the one hand there are uses of word combinations like ‘confident’, ‘ecstatic’, ‘enraptured’, ‘excited’, ‘glowing’ and ‘patient’ hope that point to a positive affective experience; but on the other hand, the corpora yielded uses of word combinations that suggest a quite different affective experience of hoping: ‘wistful’, ‘tense’, ‘uneasy’, ‘nervous’, ‘worried’, ‘anxious’, and ‘desperate’ hope. An example in context is the following: “Three children, aged between four and seven, scream as uniformed East Berlin police drag them and their mother off and push them into a lorry as they stand outside the US embassy in East Berlin, hoping desperately to get inside.” (BNC #1) I took the factual use of such word combinations to be one piece of evidence in favour of my view that hope may not only be a positive and pleasant, but also a negative and unpleasant affective experience.<sup>12</sup>

Barbara Vetter (2014) uses corpus linguistics in an ontological paper that defends her non-conditional conception of dispositions. Among the many arguments she advances, there is one that is based on a corpus analysis. It is not necessary to rehearse the details of Vetter’s position to understand that the corpus analysis is used to debunk a counter-argument levelled against Vetter’s theory of dispositions. This counter-argument is based on the linguistic intuitions of her opponents, who claim that the expression ‘disposed to’ is used in a certain way in conditional sentences structures (with ‘if’ and ‘when’); but, as Vetter observes, of the 226 pertinent occurrences of ‘disposed to’ in COCA, none supports her opponents’ view: “In the entire corpus, there is not a single example of ‘disposed to’ being used to ascribe to a concrete, inanimate subject a relatively permanent and intrinsic tendency to behave in certain ways—a disposition in the philosophers’ sense” (Vetter 2014: 16). Based on her findings, Vetter argues that her opponents’ intuitions do not constitute evidence, because the expression ‘disposed to’ in the construction that they have pointed to is a technical term. And what’s more, it is

---

11 In German: ‘vergebliche’, ‘unrealistische’, ‘überzogene’, ‘naïve’, ‘kindische’, ‘kriminelle’, ‘nährische’, and ‘unmoralische Hoffnung’. The complete list can be found in Bluhm (2012: 188f).

12 For the full argument, cf. Bluhm (2012: 139–186).

a technical term rooted in the theory that her opponents have expounded. Their counter-argument, Vetter argues, is, therefore, begging the question.

Aurélie Herbelot (Ms.) introduces some ways of using corpora in philosophy that differ from the ones I have introduced so far. They rely on computational methods based on distributional semantics, as opposed to the more pedestrian approaches I have described so far. Distributional semantics (roughly speaking) views the meaning of words as their distribution, that is, the linguistic contexts they are associated with. Suitably represented in a mathematical model, distributions can be calculated from large corpora. Herbelot introduces three ways in which this can be put to use in philosophy. In discourse analysis, it can help identify social construction through the analysis of language patterns in large corpora. Thus, Herbelot, Redecker, and Müller (2012) show that different genders are associated with different words and word fields. Distributional analyses can also be used in the history of ideas. For example, they can assist in the analysis of the use of important concepts by specific authors. And they may be used to experimentally evaluate philosophical theories by testing their formalisation on ordinary language corpora—on the condition that the theories can be taken to predict certain language patterns.

Although I have so far presented only a very limited number of examples for the use of corpus analysis in philosophy, it should have become apparent that there is wide scope for the data contained in corpora to be accessed and used for research purposes. For example, most of the examples I have discussed rely on the individual interpretation of corpus query results, but there is also the more automatized approach to corpus analyses suggested by Herbelot, which saves much of the effort involved in the extensive examination of search results. I shall say more about such differences in approach in the final section of this paper. First, I would like to discuss two other methodological options.

## **5. A similar option: Internet search engines**

Sometimes general internet search engines are used for much the same purposes as the ones I have been discussing. In recent times, it has become somewhat popular among philosophers to use general internet search engines (like Google, Yahoo, Bing or, hopefully, any of the smaller, less domineering alternatives) to find out whether something is an actual use of a concept at issue. Obviously, some degree of empirical backing is sought for claims about language use. It is, I believe, fitting to introduce some published examples of this approach here.

An attempt to use an internet search engine in philosophy in much the same way as a corpus analyst would use a corpus was made by Peter Ludlow (2005), the

earliest published example of its kind that has come to my notice. Ludlow used Google to find examples for uses of the verb ‘to know’ or the noun ‘knowledge’ with some sort of linguistic modifier. His aim was to identify what he calls “L-marked” phrases of ‘know’ and ‘knowledge’; essentially, linguistic expressions that qualify the words in a certain way and in a certain thematic domain.<sup>13</sup> Thus, for example, Ludlow queried the phrases ‘known by \* standards’ or ‘with \* standards of knowledge’, yielding results like ‘known by *any objective* standard’ or ‘with *general contemporary* standards of knowledge’. Ludlow thus shows that ‘know’ and ‘knowledge’ in fact have a variety of “L-marked positions [...] for standards of justification and evidence, for subjective certainty of the report, for the reporter’s responsibility for having and defending the knowledge, the source of the knowledge, and the mode of presentation of the content of the knowledge report” (Ludlow 2005: 20). And these findings he believes to be evidence for contextualism with respect to knowledge.

René van Woudenberg (2009) takes a different approach. The question he pursues is whether we are responsible for what we believe. Based on the claim that this idea underlies much of our social life, he wants to oppose the sceptics by pointing out that it cannot, therefore, be given up, without far-reaching and serious consequences. In one substantial part of the paper, van Woudenberg uses an internet search to support his claim by showing that “[t]he use of deontological epistemic expressions is ubiquitous” (2009: 50). He explicitly states: “My procedure will be to select one deontological term at a time (‘obligation to/duty to/ought to’, ‘permission to’ and ‘right to’) and to consider various combined hits with belief, know(ledge), forgetting, and ignorance respectively” (2009: 50). And this is indeed what he does. As is to be expected when the internet is used as the database, the examples are from a wide variety of sources, ordinary and technical. Regrettably, there is no information as to how van Woudenberg narrowed down the presumably much more extensive results. At any rate, the examples that he discusses are pertinent uses of the listed deontological expressions.

A somewhat more quantitative procedure is employed by Kevin Reuter (2011). Reuter uses findings from a web search to show that—contrary to a widely held conviction in philosophy—people do in fact distinguish linguistically between apparent and real pain. This is contrary to the philosophical tenet that, with regard to pain, there is no conceptual distinction between appearance and reality. Or, more simply, to feel a pain is to have it. Reuter observes that there is a difference in the relative frequency of expressions referring to ‘having’ vs. ‘feeling’ pain,

---

13 The core of the question is “whether the lexical structure of the verb is such that it associates the verb with certain phrases that incorporate thematic roles” (Ludlow 2005: 18).

if the pain is described as strong or intense. According to his interpretation, if people speak of ‘feeling’ pain, it is to be construed as an introspective report that allows for some uncertainty, whereas to speak of ‘having’ pain is to present it as an objective fact. The empirical part of his argument is based on findings from three search engines (Google, Yahoo, Bing). Reuter compares the number of hits received for expressions with ‘feel’ vs. ‘have’, ‘pain’, and a list of four adjectives expressing a low or high degree of intensity of pain. Internet queries notoriously do not allow the deduction of reliable statistical information, among other reasons because they yield page hits instead of listing all tokens that were found. Reuter dodges this problem by comparing the rough ratio of hits for the collocation of ‘have’ and ‘feel’ in conjunction with low and high intensity of pain yielded from the different search engines. He observes that ‘feel’ and ‘have’ are used roughly with the same frequency in reference to weak pain, but that ‘have’ is used at a much higher ratio in reference to strong pain. Based on the assumption that ‘feel’ and ‘have’ can be associated with appearance and reality respectively, Reuter takes this as an argument against the philosophical precept cited above.

There are a number of reasons in favour of using general internet search engines to query for linguistic phenomena pertinent to philosophical issues. First of all, the internet is indeed mouth-watering for any empirically-minded linguist because of its sheer size. A considered estimate from 2008 calculates that 60% of what is termed the “visible internet” is in English, amounting to an estimated 3 trillion ( $3 \times 10^{12}$ ) words (Bergh & Zanchetta 2008: 313).<sup>14</sup> Given the speed with which the internet is growing, this number today should be much higher. The second major argument for using the internet as a database is the wide diversity of the texts contained in it. The third is the relative ease with which data can be compiled from the internet. And the fourth is that not only is access to all the data relatively easy, but, on top of that, it is mostly free of charge.

There are, however, some serious downsides to using the internet as a corpus.<sup>15</sup> First of all, English (or rather, some language that resembles English) is used on the internet by a large number of speakers who have limited linguistic competence because it is not their native language. The fact that some usage can be observed is thus not necessarily good evidence for its acceptability. Secondly, and more crippling, I believe, are limitations of access. Present-day internet search engines

---

14 It is not wholly clear whether the expression “visible internet” is supposed to refer to the generally accessible *code* of web pages or to the text to be read with the help of browsers minus code and markup language tags. One would suppose, from the name, that it is the second alternative.

15 Cf. Kilgarrieff (2007) and Bergh and Zanchetta (2008) for much of the following.

offer fewer search algorithms than corpus tools do. As opposed to a well-kept corpus, the internet is also not linguistically annotated and thus lacks information that is necessary for more sophisticated queries. On top of such qualitative limitations of access, there are also quantitative limitations. The number of queries on a given day with a given search engine is limited by the search engine providers. This is unproblematic for everyday use of search engines, but poses problems for extensive research. Thirdly, internet search engines do not allow the deduction of reliable statistical information, at least not without refinement of the search results. One of the reasons is that search engines return hits by source, that is, by web page, instead of by token. Another reason is that the results of queries with general internet search engines are (usually) not reproducible because they are handled by different servers on the basis of different indices of the internet—which alone suffices to make their integration into the canon of *experimental* approaches questionable, given that a core feature of experiments is repeatability.<sup>16</sup>

It is therefore not very advisable simply to use the web as a corpus (WaC). It is viable to use data extracted from the internet as raw material for building a corpus, that is, to use the “web for corpus” (WfC), as it is called in corpus linguistics. However, while using the web for building a corpus is respectable (and might be a future option), this is not what is in fact done today when philosophers type a query into a common internet search engine.

## 6. An alternative option: Questionnaires

Corpus analysis, WaC, and WfC are all attempts to overcome the limitations of individual intuition. Another option for doing so—put to sophisticated use by experimental philosophers—is to use questionnaires in controlled settings to obtain informants’ views on how a specific concept is used. This is by no means the most common objective pursued by questionnaire studies in experimental philosophy. More common is the attempt to discern intuitive commitment to *judgements* that are relevant to some philosophical issue. However, my focus is

---

16 It has to be conceded that corpus analysis is not altogether free of this problem either. When different software tools are used on one and the same corpus, they may yield different results (cf. Anthony 2013: 150f.). However, my objection stands, because the results of using a specific software tool on a specific corpus can indeed be reproduced reliably.

on approaches that take linguistic phenomena as a starting point, something that enjoys a certain prominence in experimental philosophy, too.<sup>17</sup>

For the purpose of studying linguistic phenomena, questionnaires can be used either indirectly, by asking whether certain constructions are or are not objectionable, or even directly, by asking how the informants would characterise the meaning of a specific concept (although I am not aware of examples of this somewhat simplistic approach). One thing to be said in favour of employing questionnaires is that they, too, like corpora, complement and (in part) replace the researcher's intuitions regarding the use of expressions under discussion. Another is that questionnaires allow the posing of questions regarding very specific and very infrequent uses of such expressions. This is an advantage of questionnaires over corpus analysis, given that philosophers are often interested in more subtle variations in the use of some linguistic phenomenon. Questionnaires also allow controlled variation in order to bring out subtle aspects of language use. Unless a corpus is built from a very repetitive type of text, e.g. telephone conversations for the purpose of fixing appointments, it is unlikely that it contains many such variations.

These are considerable advantages. However, I would like to emphasise two respects in which corpus analysis may be more advantageous. First, it is to be noted that questionnaires are usually given to a very limited number of test subjects and, therefore, do not necessarily solve the problem of limited active linguistic intuition. Secondly, the major flaw of this method, in my eyes, is that it draws the informants' attention to their use of the language and thereby invites answers that do not provide information on how informants *do* use a specific concept, but on how they *believe* they use the concept—or *should* use the concept.<sup>18</sup> Psychology has, of course, established means to draw the attention of subjects away from whatever it is the experiment tries to get a grip on, but I am somewhat sceptical that this will work when one is testing for variations in language use.

I would like to support this claim with the help of an example. Patricia Bruininks and Bertram Malle (2008) have published the results of an interesting and well-designed questionnaire study on hope. One of their observations was that test subjects associated more important objects with 'hope' than they did with 'optimism', 'desire', 'wanting', and 'wishing' (cf. Bruininks and Malle 2008: 348f). However, this is not confirmed by a corpus analysis. Quite the opposite, 'hope' is commonly used in cases of tepid hope and hope for trivial things, and no association of the kind

---

17 In fact, the earliest pieces of experimental philosophy *avant la lettre* that I am aware of, Naess (1953) and (1960), are concerned (although not exclusively) with meaning.

18 In this context we should also note, if only in passing, that the construction of a questionnaire may be biased and thus bias the answers of the subjects.

observed by Bruininks and Malle is evident when we look at the use of the word in unprompted contexts. I believe that their result is due to a biasing effect. There is a powerful ideology of hope. People seem to think that hope *should* be for something important, where ‘important’ may be read subjectively as ‘something one attaches great importance to’ or objectively as ‘something that is indeed worth attaching great importance to’. If subjects are conscious that they are being asked about hope, they will, I believe, tend to answer questionnaires in a way that conforms to their normative ideas about hope.

A similar concern may be voiced with respect to empirical research on causal reasoning.<sup>19</sup> Since insights into causal reasoning may be gained from the study of causal argument, one strain of this research focusses on the study of real-life arguments; but as Deanna Kuhn observes, “researchers studying causal reasoning skills in adults have typically based their conclusions on studies of a narrow segment of the adult population in a specific context—college students in laboratory settings performing complex paper-and-pencil tasks” (Kuhn 2007: 44). The main advantage of structured interviews in lab-style settings is that they, too, allow a maximum of control and can be efficiently used to assess behaviour in response to minimally varied conditions; but even if Kuhn’s critique regarding the choice of subjects is heeded and a more diverse sample of subjects is recruited, we may still worry about the lab-context. Its artificiality might influence the subjects’ responses, not least by priming their attention to aspects of the situation that are relevant to the issue under inquiry.

One way to alleviate such worries is to turn to corpus analysis. Again, not to supplant, but to supplement other empirical methods. While there is some impressive work on causal arguments by linguists,<sup>20</sup> philosophers so far have made no use of corpus analysis in their study of causal arguments. This is somewhat surprising, because argumentation is a frequent linguistic activity and causal reasoning in turn is frequent in arguments. Linguistic data for the study of causal reasoning is therefore readily available in corpora. As a matter of fact, there is such a wide variety of linguistic markers for causality, *viz.* causal connectives, verbs, adjectives, and prepositions,<sup>21</sup> that the challenge is not to find data but not to be overwhelmed by its wealth. As I have argued above, data from corpora can be considered to be unbiased (on the condition that it has been recorded and

---

19 Cf. Hahn, Zenker, and Bluhm (forthcoming).

20 Most notably, Oestermeier and Hesse (2000).

21 Cf. Altenberg (1984); Khoo, Chan, and Niu (2002); and Diessel and Hetterle (2011).

compiled independently), and given the right source of data, we can observe argumentation in comparatively natural environments.

## 7. Ups and downs of corpus analysis

It is quite obvious that there are some disadvantages and limitations to corpus analysis as a method for philosophy. Most importantly, corpus analysis is not suitable for addressing many philosophical questions, including questions with a strong empirical component. Language is *not* always relevant. And even if it is, corpus analysis may not always be the tool of choice. For some purposes (e.g. subtle variations on some linguistic theme), questionnaire studies have clear advantages. It is also worth pointing out that corpora provide textual context, but a limited amount of other contexts. For example, if they contain only language recorded in writing, they (usually) lack information on non-verbal linguistic phenomena (e.g. prosody and breaks), para-linguistic phenomena (e.g. gestures, facial expressions, and body language), and non-linguistic context. Another limit to corpus analysis is that the evidence it provides only has so much weight. The fact that some expression *cannot* be found in a corpus does not prove anything at all; and the fact that it *can* be found does not prove that it is well-formed in the language. If you like an aphorism: the advantage of corpora is that they give evidence of actual instead of merely possible uses of language; the disadvantage is that they give merely evidence of actual uses of language and not of (all) possible ones. Given all these limitations, it is also relevant to admit that analysing a concept with the help of a corpus is a lot (!) of work, at least if it is done by hand. Such efforts are only to be undertaken if they promise worthwhile results.

Still, there are a number of benefits of using corpora—again presuming that linguistic surface phenomena matter. Most importantly, they provide independent data, removing many if not all biasing effects. On the basis of this data, hypotheses can be formulated; and hypotheses and conclusions can be confirmed or falsified with it. The data can also be used to exemplify or illustrate specific usages of interest. I think it is important to stress that one of the rationales for using corpora is that they can provide positive evidence for claims about the use of a given expression. Wherever the use of ordinary language has some argumentative force, corpus analyses, thus, may provide positive reasons for philosophical claims. I think it is worth underlining this point because my impression is that one of the predominant methods in philosophy is simply to claim something and then see whether anyone comes along and disproves the claim. That is slightly odd. I am not saying that it is always wrong, but I would hold that being able to give positive reasons is preferable.

There are also two somewhat more incidental benefits of using corpora I would like to mention: one is that the contexts in which the queried expressions are found give insights into the variety of real life situations in which the phenomenon referred to by the expressions occurs. Another is that they often provide inspiration or excellent raw material for thought-experiments with regard to the concept or the phenomenon in question.

Although I have given only a very limited number of examples for the use of corpus analysis in philosophy, and although I have only briefly sketched the features of corpus software, my account does give an inkling of the range of applications that corpus analysis may have in philosophy. In corpus linguistics, there are two general research strategies. *Corpus-driven* research uses corpora in an explorative fashion, with minimal hypotheses as to the linguistic forms relevant to a given research question; *corpus-based* research uses them to verify or falsify hypotheses about the use of language on the basis of available theories about linguistic forms.<sup>22</sup> Both strategies are mirrored in the different *functions* that corpus analysis may serve in philosophy. They can be used in an explorative manner to facilitate philosophical research for which linguistic phenomenon are somehow pertinent; but they can also be used more strictly to gather evidence to support or undermine philosophically relevant claims. I would also like to draw attention to the fact that there are differences with respect to the level of complexity exhibited by the queried expressions. Clearly, research cannot only focus on words, but also on expressions consisting of two or more words. It is also possible to make sentences the object of study, and even more complex patterns of language, an example of which are arguments spanning more than one sentence. And, last but not least, there is a difference between qualitative and quantitative approaches, the latter of which are especially tempting because they promise to much reduce the toil of interpretation.

The examples I have cited are all concerned with semantics, but research based on syntactic phenomena would also be possible,<sup>23</sup> just as, given the right corpus, pragmatics. And there are other research interests that corpus analysis may serve, although no attempt has been made to do so thus far. Thus, corpus analysis can be used to discern inter-linguistic and intra-linguistic similarities and differences. Experimental philosophy has posed the question whether there are differences in intuition between people of different native tongues. Another big issue is whether

---

22 Cf. Biber (2010).

23 This possibility is also mentioned by Louw (2011: 181f.), but I am not convinced by his example.

there are differences between the intuitions of experts and those of lay people. From a corpus linguistic standpoint, this may be viewed as a sociolinguistic exercise, for which evidence can be taken from comparative corpora.

Arianna Betti notes that “any methodological position must be applied consistently, and in the case of ordinary language philosophy I do not see how we can avoid resorting scrupulously to the wealth of empirical linguistic research available today” (Betti 2014: 50). Betti, admittedly, believes that the method of ordinary language philosophy yields untenable results, at least with respect to ontology; but I do not claim her as a proponent of ordinary language philosophy, only quote her for the assessment I wholeheartedly agree with: *if* one believes in the importance of ordinary language for philosophy, it is hard to see how empirical research is to be avoided. This is also true with respect to other questions that experimental philosophers have not only asked but are actually trying to answer—provided that there is a linguistic angle to them. Philosophers can decide to leave the empirical research to the linguists; but the happy news is that the accessibility of big linguistic text corpora through the internet has made corpus analysis a feasible option for non-linguists, too.

## References

- Altenberg, B. (1984). “Causal linking in spoken and written English”. In: *Studia Linguistica* 38(1): 20–69.
- Anthony, L. (2013). “A critical look at software tools in corpus linguistics”. In: *Linguistic Research* 30(2), 141–161.
- Bergenholtz, H. and B. Schaefer (1985). “Deskriptive Lexikographie”. In: L. Zgusta (ed.), *Probleme des Wörterbuchs*. Darmstadt: Wissenschaftliche Buchgesellschaft, 277–319.
- Bergh, G. and E. Zanchetta (2008). “Web linguistics”. In: Lüdeling, A. and M. Kytö (eds.), *Corpus Linguistics: An International Handbook*. Vol. 1. Berlin: de Gruyter, 309–27.
- Betti, A. (2014). “The naming of facts and the methodology of language-based metaphysics”. In: Reboul, A. (ed.), *Mind, Values, and Metaphysics: Philosophical Essays in Honor of Kevin Mulligan*. Vol. 1. Cham et al.: Springer, 35–62.
- Blumh, R. (2012). *Selbsttäuscherische Hoffnung: Eine sprachanalytische Annäherung*. Münster: mentis.
- BNC #1 = s.n. In: *The Independent*, 06.10.1989, 161.
- Bruininks, P. and B. F. Malle (2005). “Distinguishing hope from optimism and related affective states”. In: *Motivation and Emotion* 29(4): 327–55.

- COCA #1 = Hunnicutt, E. (2003). "Mittel Europa". In: *Literary Review* 46(3), 497–510; here: 497.
- Diessel, H. and K. Hetterle (2011). "Causal clauses: A crosslinguistic investigation of their structure, meaning, and use". In: Siemund, P. (ed.), *Linguistic Universals and Language Variation*. Berlin: Mouton de Gruyter, 23–54.
- Hahn, U., F. Zenker and R. Bluhm (forthcoming). "Causal argument". In: Waldmann, M. R. (ed.), *Oxford Handbook of Causal Reasoning*. New York: Oxford University Press.
- Herbelot, A. (Ms.). "Distributional semantics for philosophy". Paper contributed to the workshop "Empirical Methods of Linguistics in Philosophy", Dortmund, March 2014.
- Herbelot, A., von Redecker, E. and Müller, J. (2012). "Distributional techniques for philosophical enquiry". In: Zervanou, K.A. and A.P.J. van den Bosch (eds.), *Proceedings of the 6<sup>th</sup> EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Stroudsburg, PA: Association for Computational Linguistics, 45–54.
- Hoey, M. (2009). "Corpus linguistics and word meaning". In: Lüdeling, A. and M. Kytö (eds.), *Corpus Linguistics: An International Handbook*. Vol. 2. Berlin: de Gruyter, 972–87.
- Hundt, M. (2008). "Text corpora". In: Lüdeling, A. and M. Kytö (eds.), *Corpus Linguistics: An International Handbook*. Vol. 1. Berlin: de Gruyter, 168–87.
- Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London: Longman.
- Khoo, C., S. Chan and Y. Niu (2002). "The many facets of the cause-effect relation". In: Green, R., C.A. Bean and S. H. Myaeng (eds.), *The Semantics of Relationships: An Interdisciplinary Perspective*. Dordrecht: Springer Netherlands, 51–70.
- Kilgarriff, A. (2007). "Googleology is bad science". In: *Computational Linguistics* 1(1): 1–5.
- Kilgarriff, A. and G. Greffenstein (2003). "'Introduction' to 'the Web as Corpus'". In: *Computational Linguistics* 29(3): 333–47.
- Kuhn, D. (2007). "Jumping to Conclusions: Can People be counted on to make sound judgments?" In: *Scientific American Mind* February/March, 44–51.
- Lee, D. Y. W. (2010). "What corpora are available?" In: McCarthy, M. and A. O'Keeffe (eds.), *Corpus Linguistics*. London and New York: Routledge, 107–21.
- Leech, G., P. Rayson and A. Wilson (2001). *Word Frequencies in Written and Spoken English*. London: Longman.
- Louw, B. (2011) "Philosophical and literary concerns in Corpus Linguistics." In: Viana, V., S. Zyngier and G. Barnbrook (eds.), *Perspectives on Corpus Linguistics*. Amsterdam: John Benjamins, 171–196.

- Ludlow, P. (2005). "Contextualism and the New Linguistic Turn in Epistemology". In: Preyer, G. and G. Peter (eds.), *Contextualism in Philosophy: Knowledge, Meaning, and Truth*. Oxford: Clarendon, 11–50.
- McEnery, T. and Hardie, A. (2012). *Corpus Linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.
- Naess, A. (1953). "An empirical study of the expressions 'true', 'perfectly certain' and 'extremely probable'". In: *Avhandlingar utgitt av Det Norske Videnskaps-Akademi i Oslo II. Historisk-Filosofisk Klasse*, 1–41.
- Naess, A. (1960). "Typology of questionnaires Adopted to the study of expressions with closely related meanings". In: *Synthese* 12(4), 481–494.
- Oestermeier, U. and F. W. Hesse (2000). "Verbal and visual causal arguments". In: *Cognition* 75(1), 65–104.
- O'Neill, E. and E. Machery (2014). "Experimental Philosophy: What is it good for?" In: E. O'Neill and E. Machery (eds.), *Current Controversies in Experimental Philosophy*. New York: Routledge, vii–xviii.
- Perkuhn, R. and C. Belica (2006). "Korpuslinguistik – das unbekannte Wesen". In: *Sprachreport* 1/2006: 2–8.
- Reuter, Kevin (2011). "Distinguishing the appearance from the reality of pain". In: *Journal of Consciousness Studies* 18(9–10), 94–109.
- Schütze, C. T. (2010). "Data and evidence". In: Brown, K., A. Barber and R. J. Stainton (eds.), *Concise Encyclopedia of Philosophy of Language and Linguistics*. Amsterdam et al.: Elsevier, 117–123.
- Vetter, B. (2014). "Dispositions without conditionals". In: *Mind* 123(489): 129–56.
- Wiegand, H. E. (1989). "Die lexikographische Definition im allgemeinen einsprachigen Wörterbuch". In: Hausmann, F. J. et al. (eds.), *Dictionaries: An International Encyclopedia of Lexicography*. Vol. 1. Berlin: de Gruyter, 530–88.
- Woudenberg, R. van (2009). "Responsible belief and our social institutions". In: *Philosophy* 84: 47–73.
- Xiao, R. (2008). "Well-known and influential corpora". In: Lüdeling, A. and M. Kytö (eds.), *Corpus Linguistics: An International Handbook*. Vol. 1. Berlin: de Gruyter, 383–457.