Experimenting with corpora: preliminary results from a first replication of the Knobe effect using analysis of textual data

Louis Chartrand

May 15, 2017

Introduction

Experimental philosophy has developed as a way of empirically studying philosophical intuitions, be it as a way determine their content, as a means to uncover their epistemic role, or as a tool to put their reliability and validity into question (Knobe and Nichols 2008). To do so, it has mostly drawn from the methods of cognitive and social psychology; however, other methods, such as corpus methods, have been overlooked.

Corpus methods have been developed in several disciplines as ways to study linguistic and discourse phenomena. In philosophy, the research program of computer-assisted conceptual analysis of texts (or CACAT. Cf. J.-G. Meunier and Forest 2008; Chartier et al. 2008; Sainte-Marie et al. 2011) stands out, with aims similar to those that drive experimental philosophy, among which the project of developing a more empirically-informed conceptual analysis.

In this paper, I propose that the methods of CACAT can indeed be of use to projects in experimental philosophy. I argue, furthermore, that not only are they well suited for uses within the context of discovery where it has mostly developed, but that they also have the means to validate certain hypotheses. To defend this claim, I present as a proof of concept a replication of Knobe's (2003) study, where the first evidence for the side effect effect was presented.

Experiments

The side effect effect arises from a concern for identifying the features that are relevant in judging whether the consequences of an action are intentional or not. Knobe believes that, according to ordinary use of the concept, whether a side effect is helpful or harmful has an effect on whether we think it is intentional or not. In order to validate this claim, he sets up an experiment where he submits participants to scenarios and asks them to weigh in on the agent's intentionality. His claim is thus vindicated by the fact that participants judged that the agent had intentionally caused the harmful side effects, but not the positive ones.

In a first section, I analyze Knobe's study to make the relationship between the research question and the experimental protocol explicit. In so doing, I identify four key components of Knobe's question: the dependent variable (intention), the independent variables (goodness/badness), the scope (side effect) and a controlled variable (foreknowledge); and I illustrate how they shaped the scenarios and the questions that were submitted to the participants.

Drawing from this analysis, I then design two experiments to address the same question, but using corpus-based methods.

In the first experiment, I use a pretrained set of word embeddings trained on the Google News corpus (Mikolov et al. 2013), which contains 100 billion words extracted from English language news article. This dataset acts as a model for the semantics of word usage in ordinary language. Knobe's hypothesis is then translated into predictions about the relatedness relations between words corresponding to the key components of the research question. Observations on the set of word embeddings follow those predictions even when some parameters are changed, thus vindicating the side effect effect hypothesis, but without measures of statistical confidence.

In the second experiment, I seek to confirm the results obtained with word embeddings, but with methods that afford confidence scores (i.e. *p*-values). To this effect, I employ methods that rely on direct observation of co-occurrence within textual contexts (J. G. Meunier, Biskri, and Forest 2005; Chartrand, Cheung, and Bouguessa forthcoming), along with sentiment analysis (Thelwall et al. 2010). Observations were made on a relevant subset of the NOW corpus, a 4.4-billion-words corpus of online news and magazine articles. The results confirm those of the first experiment, but with a low confidence, thus failing to achieve statistical confidence[1].

The potential of corpus-based methods

While we cannot affirm the results from experiment 1 and 2 with confidence, I argue that achieving statistical significance is not critical for the aims of this proof of concept if we can argue that it remains within our grasp. The results from experiment 1, while they cannot be interpreted as hard evidence in the absence

[1] This new, massive corpus brought about unforeseen technical difficulties that couldn't be solved before the due date. Once the analysis will have been completed, I will probably be able to report results that are statistically significant, and modify my presentation or poster accordingly.

of relevant statistical tests, suggest that it might be fruitful to analyze word embedding models for semantic analysis. Therefore, developing statistical tests for this purpose would be a fruitful endeavour. The results from experiment 2, on the other hand, would probably achieve significance given a larger subcorpus and/or more refined analysis tools (e.g. Chartrand, Cheung, and Bouguessa forthcoming).

The contribution of our proof of concept is thus threefold. Firstly, it demonstrates how a research question can be analyzed and translated in terms that can be modelled and tested using CACAT experimental protocols. Secondly, it demonstrates how these protocols can effectively be implemented on corpora to produce meaningful results. And thirdly, it generates results that are coherent with the results obtained from previous studies on the side effect effect. These contributions, I argue, constitute a significant headway towards a *bona fide* CACAT validation methodology.

Finally, I highlight three potential contributions that CACAT can bring to experimental philosophy. Firstly, while corpora are not cheap, once acquired, they can constitute effective sandboxes for exploring the semantics of philosophical concepts. As such, they could help philosophers discover unexpected aspects of philosophical concepts. Secondly, CACAT brings new sources of data to experimental philosophy. In particular, while corpora are unlikely to solve the socio-cultural diversity problems raised by Weinberg, Nichols, and Stich (2001), they could enable us to observe linguistic and conceptual behaviour from groups of people which are hard to reach, but which have provided us with analyzable corpora. Thirdly, where experimental philosophy can only make a snapshot of the present, CACAT opens up the possibility of studying the past, or even studying the evolution of concepts or other linguistic entities across time. As such, echoing the wish formulated by Knobe and Nichols (2008) to reestablish the place of history of philosophy in analytic philosophy, CACAT could bring a new life to genealogical analysis of concepts (S. Haslanger and Haslanger 2012).

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, Jackie Cheung, and Mohamed Bouguessa. forthcoming. "Detecting Large Concept Extensions for Conceptual Analysis." New York: 13th International Conference on Machine Learning and Data Mining (MLDM 2017).

Haslanger, Sally, and Sally Anne Haslanger. 2012. *Resisting Reality: Social Construction and Social Critique.* Oxford University Press.

Knobe, Joshua. 2003. "Intentional Action and Side Effects in Ordinary Language." *Analysis* 63 (279). Wiley Online Library: 190–94.

Knobe, Joshua, and Shaun Nichols. 2008. "An Experimental Philosophy Manifesto." *Experimental Philosophy* 1. New York: Oxford University Press: 3–14.

Meunier, Jean Guy, Ismail Biskri, and Dominic Forest. 2005. “Classification and Categorization in Computer Assisted Reading and Analysis of Texts.” Elsevier.

Meunier, Jean-Guy, and Dominic Forest. 2008. “Computer Assisted Conceptual Analysis of Text: the Concept of Mind in the Collected Papers of CS Peirce.” *Digital Humanities 2008*, 163.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” *CoRR* abs/1301.3781. http://arxiv.org/abs/1301.3781.

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). ALLC: 329–34.

Thelwall, Mike, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. “Sentiment Strength Detection in Short Informal Text.” *Journal of the American Society for Information Science and Technology* 61 (12). Wiley Online Library: 2544–58.

Weinberg, Jonathan M., Shaun Nichols, and Stephen Stich. 2001. “Normativity and Epistemic Intuitions.” *Philosophical Topics* 29 (1/2). University of Arkansas Press: 429–60.