Distant Reading and Data-Driven Research in the History of Philosophy

DR2

Working Papers 1

Accademia university press

Distant Reading and Data-Driven Research in the History of Philosophy

DR2

Working Papers 1

ISSN 2785-4310

La pubblicazione del presente volume è stata realizzata con il contributo dell'Università degli Studi di Torino, Dipartimento di Filosofia e Scienze dell'Educazione

ISBN 9791280136800

dr2blog.hcommons.org
www.aAccademia.it/DR2-1-2021

prima edizione: dicembre 2021

aAccademia university press

Contents

On some challenges posed by corpus-based research in the history of ideas

Enrico Pasini

Abstract. The paper briefly focuses on some peculiarities of corpus-based research when we move from linguistics to the history of ideas (and of concepts, doctrines and arguments). Building corpora for this kind of research poses specific challenges in terms of collection strategies, design decisions, and annotation.

1. *Introduction*

The history of ideas, or history of concepts (also known as *Begriffsgeschichte*) can be seen either as a field of historical research, or as a set of historical methodologies and approaches, in both cases dealing with the historical semantics of certain terms and expressions: it studies how they arise and are transmitted, related, and modified over time. It intersects in varying ways intellectual history, as well as the history of philosophy, of medicine, of science, of literature. In this broad research setting, a growing interest for computational techniques applied to textual sources has surfaced in recent times, often in the form of corpus-driven, or corpus-based research. This approach can easily fall prey to the lure of seemingly magical techniques: as it was for microscopes and telescopes, just put your eye on them and see new things. And it can be witness to the limits, indeed the perils of claustrophobic corpora, when it confines itself to some present-time equivalent of 1990s OCR-ed collected works of an author.

The purpose of this contribution is to discuss some requirements of corpus-based research in the history of ideas as far as corpora themselves are concerned.

2. *Corpora: a many-sided challenge*

What's in a corpus? "There are many ways to define a corpus ... but there is an increasing consensus that a corpus is a collection of (1) machine readable (2) authentic texts (…) which is (3) sampled to be (4) representative of a particular language or language variety." (McEnery et al. 2006, 5)

DR2 Working Papers 1, 2021, pp. 9-17
ISSN 2785-4310

ISBN 9791280136800

The driving forces for the building of corpora have been, quite obviously, corpus linguistics and natural language processing. This has brought to remarkable results and even to a considerable level of automation in corpora development. In this context, corpora are used to represent a language, or a variety of a language, or a subset of a language (as in the study of English for specific purposes), possibly across languages. Important publicly available corpora are, indeed, specifically built to this purpose in nearly every case.

By contrast, building corpora for the use of the history of ideas and concepts, of doctrines and arguments, presents a many-sided challenge, in terms both of collection strategies and of design decisions.

2.1. *Content*

The challenge is partly a sheer matter of content—if historical and disciplinary criteria are to take the place of linguistic criteria, the relevant texts—in adequate size and quantity—are simply not there to be used. The shortcomings of existing sources and the difficulties in the production of suitable corpora for a "computational turn" in the history of philosophy were discussed by Arianna Betti and her co-authors in Betti et al. (2017, 379). There, the necessity of such resources was strongly advocated: "it is crucial to be able to build high-quality, easily and freely accessible corpora in a sustainable format composed from multi-language, multi-script books from different historical periods." Yet, as for this there is no great difference between the situation in 2017 and in 2020.

Of course, as in times of drought the byline can be: do with the sources you have—curate, document, and improve. This, in fact, would also be the mantra with more ambitious projects concerning more ambitious corpora. But decisions must anyway be taken with regard to quality, granularity, saturation (in the phase of the choice of texts), and with regard to correction and annotation (in the phase of curation).

1. In the design of suitable corpora that would collect a sufficient amount of such content, there are some problems to be considered, e.g. whether one should privilege historical editions over the so-called *Ausgabe letzter Hand*, or the best available critical edition, or vice-versa. The obvious risk is to end up with an overall corpus that would be sort-of representative of a *perennis philosophia*, but not necessarily representative of textual production over time (from both the synchronical and the diachronical point of view, if equivalent or comparable slices could be singled out).

2. Adding interdisciplinarity as a necessary, unavoidable feature of the objects of the history of ideas (Albertone and Pasini 2014), one should go beyond the works of the philosophers, to include such different text types as, on different axes: scientific, philosophical, medical, religious; narrative, descriptive, argumentative; high/low; private, public. It seems, in other words, that such a corpus should aim to inclusiveness. It should ideally be comprehensive enough that no text could be considered as "rogue" and as a menace to the homogeneity of the corpus.[1] Large, genre-balanced corpora relevant to the history of concepts and ideas would be, mimicking an over-famous definition by Sinclair (1991, 71), one or more collections of historically given, published and/or unpublished texts, chosen to characterise a state or variety of a cultural configuration through certain specific conceptual and argumentative features of their content. Such a collection of texts would be used both as a means of verifying hypotheses about historical sets of concepts, and to extract experimentally the features of those historical sets.

3. It may sound obvious, but also size,[2] and not only composition and quality, must be taken into consideration. This poses, if not a theoretic, some practical challenges—such that call if nothing else for the usual words of wisdom: "The construction of a large scale corpus is a hard task" (Li et al. 2007, 56), and "a measure of compromise is often necessary" (Sinclair 2005, 79). As far as the history of ideas is concerned, "authors have been writing for millennia"! (Michel et al. 2011, 177).

4. Moreover, multi-linguism requires to produce either comparable corpora or parallel corpora, the latter being corpora that contain source texts and their translations (McEnery and Xiao 2007, 2), while the former are not translations of each other, with "similar balance and representativeness (…), e.g. the same proportions of the texts of the same genres in the same domains in a range of different languages in the same sampling period" (McEnery and Xiao 2007, 3). Important sets of historical translations (e.g. into English or between Latin and vernaculars in the 17th-18th century) could be used to this purpose. Yet most of the textual production in question allows only for comparability. But a suitable comparability metric[3] would be needed to estimate the quality of a corpus built on the same topics in different languages.

[1] "A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided" (Sinclair 2005, 14).
[2] It would easily represent "big data" in the sense of De Mauro et al. 2016.
[3] See for instance Su and Babych 2012.

2.2. *Scope and aim*

When speaking of historical corpora, it is important to consider the difference between "track[ing] changes in language evolution" (McEnery et al. 2006, 65) and tracking changes in the evolution of concepts, arguments, and theories. And although the makers of Google n-grams wrote that "Culturomics has profound consequences for the study of language, lexicography, and grammar" (Michel et al. 2011, 178), there is also difference from their tracking of "cultural trends" (Michel et al. 2011, 176).

Admittedly, from the point of view of computational linguistic the analysis of corpora, as we read in Mark Davies' description of the BYU collection, can be presented as useful in "gaining insight into culture; for example what is said about different concepts over time and in different countries" (Davies, n.d.). In this way of looking, concepts are given objects about which something may be said: to discourse on concepts, or on insects, or on other constituents of the universe, is an essentially linguistic phenomenon that can be studied comparatively with standard techniques.

Yet concepts, in all but strongly Platonic approaches, and especially from the point of view of intellectual history, or of the history of philosophical and scientific thought, *do not exist someplace and are not spoken about*: on the contrary, it is precisely what has been said with conceptual words that makes concepts, or makes it possible that conceptual formations (conceptual patterns, constellations, frameworks, networks, and hierarchies) are studied in their structure and development.[4]

2.3. *Annotation/detection*

There is a manifest difference between raw and annotated corpora, between corpora for hypotheses formation and corpora for testing hypotheses. At least two questions arise. Would we have representative sub-corpora from which to extract categories, lexica, and explicit and implicit hierarchies, and would we have the time to loop between experiment and analysis, if the main corpus be gigantic? Do we have appropriate techniques, or can we appropriate techniques: e.g., adapt named entity recognition to concepts, considered as named individual entities, or borrow from unsupervised ontology

[4] This is not a circularity of the bad kind. It is felicitously mirrored in the suggestion to "proceed in cycles" when trying to achieve "representativeness" in linguistic corpus design as discussed by Biber 1993; something similar in Plappert's recurring cycles of induction on trigrams (Plappert 2017).

detection,[5] or look for "introducers":[6] although it might be difficult to use automatic annotation software, we might pick out typical conceptual relations (e.g. 'is a kind of').

Such considerations are in a way inevitable and, moreover, they make all the more necessary an effort to define some kind of *standardised conceptual annotation*, be it a priori or a posteriori, to ensure comparability between different annotated corpora that might be developed in parallel.

This again poses the practical problem, not new, of the trade-off between precision and time: automatic tagging vs. crowd-sourced tagging vs. expert tagging. But a more radical question poses itself: is it possible to tag without a theory, or with but the suspension of a theory? Do we have categories or domains broad enough, and still significant?

2.4. *Semantics, and a critical point*

Corpora based on the kind of broad historical sources that would be needed are rare. A good example, and the only one among those listed at BYU, is the annotated version of *Early English Books Online* (EEBO). The corpus was created as part of the SAMUELS project (Semantic Annotation and Mark-Up for Enhancing Lexical Searches). The corpus is annotated semantically, and searchable by semantic tags: more than 8,000 different semantic categories, based on the University of Glasgow's Thesaurus (Historical Thesaurus of English, n.d.). So, again, *it presupposes the pre-existence of the very concepts that are tagged*. It presupposes for instance theories—at least a model—of the basic emotions, and of the corresponding moods, of which a verbal expression has been coded as a hypernym. To this model are reduced the very varieties of classification and conceptual hierarchy, the history of which we would like to study.

We see here clearly, insofar as the history of concepts departs from a special lexicography, how specific theoretic and practical problems arise. Could hypernym hierarchies be created automatically for the history of concepts? It is easy to appreciate that from the point of view of the history of concepts, hyponym/hypernym relations are not the same and do not work in the same way as in general language, not only because of the genetic and diachronic aspects, but because of the different role of logical constraints, on the one

[5] See Toledo-Alvarado and Martínez-Luna 2012.
[6] Like 'north of', 'next to' in geographical information retrieval (Sallaberry 2013).

hand, and of the import of competition between theories and models, on the other hand.

3. *Argument mining*

Finally, argument mining should be mentioned in relation to corpora and the history of (philosophical, medical, scientific) ideas—and in this case quite specialized corpora might be intended. Argument detection would be a powerful technique in the history of ideas, as for both philosophic and scientific thought, and work has been done, in fact, on raw corpora, looking for presence and frequency of argument markers (for instance 'consequently', 'since', and the like),[7] but with results that are not decisive. In the DR2 group, we are considering instead the production of annotated corpora both of historical texts, and of paradigmatic texts (e.g. handbooks of logic) to train detection procedures and to use as test corpora.[8] We have begun to design the annotation process, in terms of arguments, spans of text surrounding the argument, granularity of annotation, and we expect, after the pandemic-caused interruption, to be working in collaboration with the instructors of logic and philosophy of science in our university and their students.

Acknowledgments

This paper is largely based on ongoing discussions in the Research group on Distant Reading and Data-driven Research in the History of Philosophy (DR2, https://dr2blog.hcommons.org/). I'd like to thank the members of the research group for providing the subject matter and much discussion for this paper, Alessandro Mazzei for convincing me to set down its first version, and the participants to a session on corpora during the 2019 DR2 Conference for a stimulating discussion.

[7] Monaco and Puente-Castelo (2013) used e.g. 'if' and other conjunctions to categorize different types of conditionals in a corpus of English scientific writings.

[8] Roughly similarly to the use of a test corpus described by Bosc et al. 2016a; see also Bosc et al. 2016b for a completely different context. A different approach (study of structural features in argumentation to unearth argumentative sections, like for instance in Stab and Gurevych 2017) deserves attention but seems less suited for working on historical texts.

References

Albertone, Manuela, and Enrico Pasini. 2014. "Introduction." *History of European Ideas* 40 (Special issue: Interdisciplinary History of Ideas): 451–456. https://doi.org:10.1080/01916599.2013.826433.

Aman, Saima, and Stan Szpakowicz. 2007. "Identifying Expressions of Emotion in Text." In *Text, Speech and Dialogue*, edited by Václav Matoušek and Pavel Mautner: 196–205. Berlin: Springer. https://doi.org/10.1007/978-3-540-74628-7_27.

Bendaoud, Rokia, Mohamed Rouane-Hacene, Yannick Toussaint, Bertrand Delecroix, and Amedeo Napoli. 2007. "Text-Based Ontology Construction Using Relational Concept Analysis." In *International Workshop on Ontology Dynamics – IWOD 2007*. Innsbruck. https://hal.inria.fr/inria-00167681.

Betti, Arianna, Martin Reynaert, and Hein van den Berg. 2017. "@PhilosTEI: Building Corpora for Philosophers." In *CLARIN in the Low Countries*, edited by Jan Odijk and Arjan van Hessen, 379–392. London: Ubiquity Press. https://www.ubiquitypress.com/site/chapters/10.5334/bbi.32/.

Biber, Douglas. 1993. "Representativeness in Corpus Design." *Literary and Linguistic Computing* 8 (4): 243–257. https://doi.org/10.1093/llc/8.4.243.

Bosc, Tom, Elena Cabrio, and Serena Villata. 2016a. "DART: A Dataset of Arguments and Their Relations on Twitter." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, edited by Nicoletta Calzolari, 1258–1263. Paris: ERLA.

———. 2016b. "Tweeties Squabbling: Positive and Negative Results in Applying Argument Mining on Social Media." In *Computational Models of Argument: Proceedings of COMMA 2016*, 21–32. Amsterdam: IOS Press.

Civiliene, Gabriele Salciute. 2020. "Between Surface and Depth: Towards Embodied Ontologies of Text Computing across Languages." *Interdisciplinary Science Reviews* 45 (2): 1–24. https://doi.org/10.1080/03080188.2020.1764800.

Davies, Mark. 2010. "The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English." *Literary and Linguistic Computing* 25 (4): 447–464. https://doi.org/10.1093/llc/fqq018.

———. 2017. "Early English Books Online (EEBO)." Accessed 24 October 2020. https://corpus.byu.edu/eebo/.

———. n.d. "BYU Corpora: Overview." Accessed 24 October 2020. https://corpus.byu.edu/overview.asp.

Debole, Franca, and Fabrizio Sebastiani. 2005. "An Analysis of the Relative Hardness of Reuters–21578 Subsets." *Journal of the American Society for Information Science and Technology* 56 (6): 584–596. https://doi.org/10.1002/asi.20147.

De Mauro, Andrea, Marco Greco, and Michele Grimaldi. 2016. "A Formal Definition of Big Data Based on Its Essential Features." *Library Review* 65 (3): 122–135. https://doi.org/10.1108/LR-06-2015-0061.

Guldi, Jo. 2018. "Critical Search: A Procedure for Guided Reading in Large-Scale Textual Corpora." *Journal of Cultural Analytics* 1 (2): 1–35. https://doi.org/10/ggd5s2.

Hall, David, Daniel Jurafsky, and Christopher D. Manning. 2008. "Studying the History of Ideas Using Topic Models." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing – EMNLP '08*, 363. Honolulu (HI): Association for Computational Linguistics. https://doi.org/10/bzz22p.

Iliev, Rumen, and Robert Axelrod. 2016. "Does Causality Matter More Now? Increase in the Proportion of Causal Language in English Texts." *Psychological Science* 27 (5): 635–643. https://doi.org/10.1177/0956797616630540.

Juola, Patrick. 2013. "Using the Google N-Gram Corpus to Measure Cultural Complexity." *Literary and Linguistic Computing* 28 (4): 668–675. https://doi.org/10.1093/llc/fqt017.

Li, Peifeng, Qiaoming Zhu, Peide Qian, and Geoffrey C. Fox. 2007. "Constructing a Large Scale Text Corpus Based on the Grid and Trustworthiness." In *Text, Speech and Dialogue*, edited by Václav Matoušek and Pavel Mautner: 56–65. Berlin: Springer. https://doi.org/10.1007/978-3-540-74628-7_10.

McEnery, Anthony M., and R. Z. Xiao. 2007. "Parallel and Comparable Corpora: What Are They up To?" In *Incorporating Corpora: Translation and the Linguist*, edited by G. James and G. Anderman. Clevedon, UK: Multilingual Matters. http://www.comp.eprints.lancs.ac.uk/59/.

McEnery, Anthony M., Richard Xiao, and Yukio Tono. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. Taylor & Francis.

Michel, J.-B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014): 176–182. https://doi.org/10.1126/science.1199644.

Monaco, Leida, and Luis Puente-Castelo. 2013. "Conditionals in Eighteenth-Century Philosophy Texts: A Corpus-Based Study." Poster presented at *Corpus Linguistics 2013*. Leicester. https://www.researchgate.net/publication/291346120

Pawlicka, Urszula. 2017. "Data, Collaboration, Laboratory: Bringing Concepts from Science into Humanities Practice." *English Studies* 98 (5): 526–541. https://doi.org/10/gf2v8b.

Plappert, Garry. 2017. "Candidate Knowledge? Exploring Epistemic Claims in Scientific Writing: A Corpus-Driven Approach." *Corpora* 12 (3): 425–457. https://doi.org/10.3366/cor.2017.0127.

Sallaberry, Christian. 2013. *Geographical Information Retrieval in Textual Corpora*. John Wiley & Sons.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

———. 2005. "Corpus and Text – Basic Principles + Appendix: How to Build a

Corpus." In *Developing Linguistic Corpora: A Guide to Good Practice*, edited by Martin Wynne: 1–16, 79–83. AHDS Guides to Good Practice. Oxford: Oxbow Books.

Stab, Christian, and Iryna Gurevych. 2017. "Parsing Argumentation Structures in Persuasive Essays." *Computational Linguistics* 43 (3): 619–659. https://doi.org/10.1162/COLI_a_00295.

Su, Fangzhong, and Bogdan Babych. 2012. "Measuring Comparability of Documents in Non-Parallel Corpora for Efficient Extraction of (Semi-)Parallel Translation Equivalents". In *EACL 2012. Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation*: 10–19.

"The Historical Thesaurus of English." n.d. Glasgow: University of Glasgow. Accessed 24 October 2020. https://ht.ac.uk/.

Toledo-Alvarado, J. I., and G. L. Martínez-Luna. 2012. "Automatic Building of an Ontology from a Corpus of Text Documents Using Data Mining Tools." *Journal of Applied Research and Technology* 10: 398–404.