JADT' 18

PROCEEDINGS OF THE
14TH INTERNATIONAL CONFERENCE
ON STATISTICAL ANALYSIS OF TEXTUAL DATA

(Rome, 12-15 June 2018)

Vol. I

UniversItalia
2018

Exploring the history of American philosophy in a computer-assisted framework

Guido Bonino[1], Davide Pulizzotto[2], Paolo Tripodi[3]

[1]Università di Torino – guido.bonino@unito.it
[2] LANCI, Université du Québec à Montréal – davide.pulizzotto@gmail.com
[3]Università di Torino – paolo.tripodi@unito.it

Abstract

The aim of this paper is to check to what extent some tools for computer-assisted concept analysis can be applied to philosophical texts endowed with complex and sophisticated contents, so as to yield results that are significant not only because of the technical success of the procedures leading to the results themselves, but also because the results, though highly conjectural, are a direct contribution to the history of philosophy

Sommario

Lo scopo di questo articolo è di verificare in che misura la computer-assisted concept analysis possa essere applicata a testi filosofici di contenuto complesso e sofisticato, in modo da produrre risultati significativi non solo dal punto di vista del successo tecnico delle procedure, ma anche in quanto i risultati stessi, sebbene altamente congetturali, costituiscono un contributo diretto alla storia della filosofia.

Keywords: philosophy, history of philosophy, paradigm, necessity, idealism, Digital Humanities, Text Analysis, Computer-assisted framework

1. Computer-assisted concept analysis

The development of artificial intelligence poses a methodological challenge to the humanities. Many traditional practices in disciplines such as philosophy are increasingly integrating computer support. In particular, *Concept Analysis* (CA) has always been a common practice for philosophers and other scholars in the humanities. Thanks to the development of *Text Mining* (TM) and *Natural Language Processing* (NLP), computer-assisted text reading and analysis can provide the humanities with new tools for CA (Meunier and Forest, 2005), making it possible to analyze large textual corpora, which were previously virtually unassailable. Examples of computer-assisted analyses of large corpora in philosophy are Allard et al., 1963; McKinnon, 1973; Estève et al., 2008; Danis, 2012; Sainte-Marie et al., 2010; Le et al., 2016; Meunier and Forest, 2009; Ding, 2013; Chartrand et al., 2016; Pulizzotto et al., 2016; Slingerland et al., 2017. The use of computer-

assisted text analysis is also relevant for the distant reading approach, developed by Franco Moretti in the context of literature studies (Moretti, 2005; Moretti, 2013), but which we are convinced can be usefully extended to different fields (for the application to philosophy see the Conference "Distant Reading and Data-Driven Research in the History of Philosophy" held in Turin in 2017, http://www.filosofia.unito.it/dr2/).

The main aim of this paper is to check to what extent some tools for computer-assisted CA can be applied to texts endowed with complex and sophisticated contents, so as to yield results that are significant not only because of the technical success of the procedures leading to the results themselves, but also because the results, though highly conjectural, are a direct contribution to the humanities. Philosophy, in particular the history of philosophy, seems to be a good case to be considered, because of the sophistication of its contents. Our main purpose is that of illustrating some of the different *kinds* of work that can be done in history of philosophy with the aid of computer-assisted CA.

2. Method

2.1. The corpus

To understand how TM and NLP can assist the work in history of philosophy, some standard methods have been applied to a specific corpus, which is provided by Proquest (www.proquest.com). The corpus is a collection of 20,751 PhD dissertations in philosophy discussed in the US from 1981 to 2015. It therefore contains 20,751 *documents*: each document is a text, comprising the title and the abstract of a dissertation, which are dealt with as a single unit of analysis. The corpus also contains some metadata, such as the author of the dissertation, the year of publication, the name of the supervisor, the university, the department, and so forth. In the present paper we are not going to exploit fully the wealth of information provided by these metadata, which are certainly worth being the subject of further research. However, we will use the crucial datum of the year of publication, which allows us to assume a diachronic (that is, historical) perspective on the investigated documents.

2.2. Data preprocessing

A preliminary step consists in a set of four preprocessing operations that allow us to extract the linguistic information needed for the analysis: 1) *Part of Speech* (POS) tagging; 2) *lemmatization*; 3) *vectorization*; 4) selection of the sub-corpora responding to Keyword In Context (KWIC) criteria.

The POS tagging and the lemmatization process are performed on the basis of the *TreeTagger* algorithm described by Schmid, 1994 and 1995. This

operation consists in the annotation of each word for each document according to its morphological category. Some irrelevant categories (such as determinants, prepositions and pronouns) are eliminated. Nouns, verbs, modals, adjectives, adverbs, proper nouns and foreign words are taken into account. The lemmatization process reduces a word to his lemma, according to the correspondent POS tag. At the end of this process, we can identify 17,750 different lemmas, which are called *types*.

The mathematical modeling of each document into a vector space is called vectorization. In such a model, each document is encoded by a vector, whose coordinates correspond to the TF-IDF weighting of the words occurring in that document. This weighting function calculates the normalized frequencies of the words in each document (Salton, 1971). At the end of the process, a matrix M is built, which contains 20,571 rows corresponding to each document, and 17,750 dimensions, corresponding to the types.

Finally, three sub-corpora are created on the basis of the *KWIC* criterion. These sub-corpora correspond to the set of all the text segments in which one of these three *lexical form*, each of which convey the meaning of a concept, appears: 'necessity', 'idealism', and 'paradigm'. The three concepts have been chosen because of the considerable diversity of their statuses: 'necessity' has always been a keyword of several sub-fields of philosophy; 'idealism' refers both to a philosophical current, historically determined, and to an abstract position in philosophy; 'paradigm' entered the philosophical vocabulary in relatively recent times, mainly after the publication of Kuhn, 1962, as a technical term in the philosophy of science. We obtain a set of 719 documents for 'necessity', 450 documents for 'idealism', 975 documents for 'paradigm'.

2.3. Word-sense disambiguation process

For each sub-corpus, we identify the *semantic patterns* (usually, word co-occurrence patterns) associated to each lexical form, so as to discover the most relevant *semantic structures* of that concept. This is done by using *clustering*, a common method in *Machine Learning* for pattern recognition tasks (Aggarwal and Zhai, 2012). Clustering techniques applied to texts are based on two hypotheses: a *contiguity hypothesis* and a *cluster hypothesis*. The former states that texts belonging to the same cluster form a contiguous region that is quite clearly distinct from other regions, while the latter says that texts belonging to the same cluster have similar semantic content (Manning et al., 2009, p. 289 and 350). For our purposes, clustering is an instrument for semantic disambiguation. In our experiment, we use the *K-means* algorithm (Jain, 2010, p. 50), a widely employed algorithm for *Word-Sense Disambiguation* tasks (Pal and Saha, 2015).

The main parameter that needs to be tuned in the K-means algorithm is the *k*

parameter, which determines the number of *centroids* to be initialized. Each execution of the K-means algorithm generates a partition P_k having a number of clusters equal to *k*. Since each centroid is the "center vector" of each cluster, it can also be used to identify the most "prototypical" documents in a given cluster. To complete this operation, a tool generally used to select relevant documents in *Information Retrieval* is employed, that is, the *cosine computation* among a *query* vector and a group of "document vectors" (Manning et al., 2009). In this context, each centroid of a P_k partition can be used as a query in order to identify documents with a higher cosine value.

Clustering has first been applied *synchronously* on the S_i matrices with $k = \{2, 3, 4, …, 50\}$, thus obtaining the most recurring semantic patterns; then it has been applied *diachronically*, dividing each matrix into three different periods (1981-1993, 1994-2003, 2004-2015) in order to obtain sets of documents with similar cardinality. On each sub-matrix of S_i several clusterings with $k = \{2, 3, 4, ..., 50\}$ were performed, in order to identify the temporal evolution of the most important semantic patterns associated to the three concepts under study. For each generated P_k partition, we also perform the cosine computation in order to obtain a set of the most relevant PhD dissertations belonging to each cluster.

3. Analyses

In this section, we are going to present three analyses, focusing on three different concepts: paradigm, necessity and idealism. Each case illustrates a different kind of historical-philosophical result.

3.1. Necessity

After exploring both synchronically and diachronically several clusters (with different k) associated to the concept of necessity, we have focused on a clustering with k=18 in the period 1981-2015 (the clusters are not significantly different from one another in the three decades). It turns out that there are at least 16 clearly distinct and philosophically interesting meanings of 'necessity': two (maybe distinct) theological notions; physical necessity; political necessity; necessity as investigated in modal logic and possible world semantics; moral necessity; necessity as opposed to freedom in debates over determinism; the necessity of historical processes; metaphysical necessity; two notions of causal necessity (attacked by Hume); the necessity of life events; logical necessity; phenomenological necessity; necessity of the Absolute (Hegel); necessity of moral duty (Kant); ancient concept of necessity; the necessity of law. In addition to these, there is also a rather big cluster in which 'necessity' seems to occur mainly with its ordinary, not strictly philosophical meaning.

If the clustering we applied to 'necessity' were extended to a large number of philosophical words (chosen in our corpus by domain experts), that would be the first step for the construction of a bottom-up vocabulary of philosophy, and ultimately of a data-driven philosophical dictionary, in which the different (though related) meanings of philosophical terms would be determined on the basis of actual use, rather than merely on the lexicographer's discernment. This lexicographic work is also an indispensable step if one wants to overcome the "concordance approach": it seems to us that this bottom-up lexicography could be a promising starting point for the construction of semantic networks.

3.2. Idealism

Unlike 'necessity', the term 'idealism' has different distributions in the decades 1981-1993, 1994-2003 and 2004-2015. We have only considered the largest clusters (> 10 documents), since for our purpose (that of reconstructing the main historical developments of American academic philosophy), isolated cases and minor tendencies are not relevant.

The evolution of some clusters over decades suggests interesting historical reflections. First, the cluster "Kant" is persistently important. In fact, it becomes more and more important, even in wider contexts, that is, in documents that are not directly devoted to Kant. This is shown by the rising trend of the cluster "Transcendental" (a term typically, but not always directly connected with Kant). Second, the cluster "Hegel" disappears in the second decade, then it reappears: is this a real phenomenon, rather than a statistical artefact? How can it be explained? Third, the cluster "Realism" disappears in the third decade: is there a relationship between the return of "Hegel" and the disappearance of "Realism"? This is not the kind of question, which comes naturally to the mind of the historian of philosophy, on the basis on his/her knowledge of well-known developments of the history of recent American philosophy. This hypothesis can be formulated only thanks to some sort of defamiliarization (*ostranenie*) with respect to the received views in history of philosophy. Yet, it seems unlikely that philosophers in the last decade gave up speaking of realism. The received view may after all be correct, that realism is more and more central in late analytic philosophy (think, for example, of the centrality of David Lewis) (Bonino and Tripodi forthcoming). Such a view is confirmed by other data, such as the number of occurrences of 'realis-' in the abstracts of the corpus. 1981-93: 373 (5,76% of 6,471); 1994-2003: 465 (6,31% of 7,361); 2004-2015: 482 (5,6% of 8,585). Thus the focus on realism is still there, in the third decade. One is therefore led to formulate an alternative hypothesis: philosophers ceased to speak of idealism *in relation to realism*: perhaps the contrast realism-

idealism has become less important than many used to think; perhaps after Dummett, realism is contrasted with anti-realism, rather than with idealism; perhaps some sort of "interference" is here produced by the presence of a further opposition, that between realism and nominalism.

The moral of this example is that clustering applied to large and conceptually sophisticated corpora allows the historians of philosophy to concoct alternative stories to account for the historical facts. This indicates that the data-driven approach can trigger the production of conjectures one would not think about. It is usually maintained that statistical techniques are useful in that they restrict the space of possible interpretations (Mitchell, 1997), but in other cases, such as the one described in this section, at least in an early phase of the hermeneutic process, in virtue of their defamiliarizing impact they can also have the opposite effect: that of broadening that same space and discovering *nouveaux observables* (Rastier, 2011).

3.2. Paradigm

This case study deals with the term 'paradigm' in the period 1981-2015. After exploring several k in the three decades, we focus on the synchronic analysis of the set of clusters with k=16. The first result that immediately stands out is that 'paradigm' occurs rather often: 995 documents, twice as many as 'idealism' (450), and considerably more than 'necessity' (719), a concept which is widely regarded as central in the recent history of Anglo-American philosophy. Using Google Ngram Viewer, and thus taking into account a generalist, non disciplinary corpus, it turns out that such a high frequency is peculiar to the philosophical discourse (the lowest value of 'necessity' is 0.0025%, which is higher than the highest value for 'paradigm', which is 0.0016%).

Why does 'paradigm' occur so frequently? On the one hand, one could find this datum not so surprising, since 'paradigm' is a technical term in the philosophy of science, introduced by Kuhn, 1962 to refer to a set of methodological and metaphysical assumptions, examples, problems and solutions, a vocabulary, which are taken for granted, in a given period of normal science, by a scientific community. On the other hand, moving from a priori considerations to the examination of the data, a partly different landscape emerges: 'paradigm' seems to be a fashionable concept, which is used in a variety of contexts as a term that is neither technical nor simply ordinary. Only in cluster 8 has the term a straightforward technical use, derived from Kuhn's philosophy of science. Each of the other clusters (1: theology, 2: music, 3: philosophy of law, 4: education; 5: nursing; 6: philosophy of religion; 7: moral philosophy; 9: bioethics, 10: spiritualism; 11: political theory; 12: self narrative; 13: theology; 14: Kant-Leibniz; 15

aesthetics; 16: philosophy and language in Wittgenstein, Heidegger etc.) does not correspond to a different meaning of the term 'paradigm', but simply to the application of the same concept to different fields. In most cases we have to do with non-technical contexts, in which 'paradigm' has neither its original grammatical meaning nor its ordinary, non-philosophical meaning (standard, exemplar). It seems to us that its meaning and use are generic and vague, rather than precise and technical; nonetheless, they evoke Kuhn: a quasi-Kuhnian vocabulary became fashionable; it entered many philosophical discourses, often more "humanistic" than "scientific" in spirit, and much less technical than the philosophy of science.

This case study expresses an especially interesting kind of result obtainable by using TM and NLP techniques to assist research in history of philosophy: it shows how the interpretation of clusters fosters the discovery of terminological fashions as opposed to genuine conceptual developments.

References

Aggarwal C.C., and Zhai C.X. (2012). "A Survey of Text Clustering Algorithms." In *Mining Text Data*, 77–128. Springer.

Allard M. et al. (1963). *Analyse conceptuelle du Coran sur carte perforées.* Mouton.

Bonino G. and Tripodi P. (eds.), *History of Late Analytic Philosophy*, special issue of "Philosophical Inquiries", forthcoming.

Chartrand L., Meunier J.-G. and Pulizzotto D. (2016). CoFiH: A heuristic for concept discovery in computer-assisted conceptual analysis. In Mayaffre D. et al. (eds.), *Proceedings of the 13th International conference on statistical analysis of textual data*, vol. I, pp. 85-95.

Danis J. (2012). *L'analyse conceptuelle de textes assistée par ordinateur (LACTAO); une expérimentation appliquée au concept d'évolution dans l'œuvre d'Henri Bergson*. Université du Québec à Montréal (http://www.archipel.uqam.ca/4641/1/M12423.pdf).

Ding X. (2013). A text mining approach to studying Matsushita's management thought. *Proceedings of the 5th International conference on informatin, process and knowledge*, pp. 36-39.

Estève R. (2008). Une approche lexicométrique de la durée bergsonienne. *Actes des journées de la linguistique de corpus*, vol. 3: 247-258.

Jain A.K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, vol. 31(8): 651-666.

Kuhn T.S. (1962). *The structure of scientific revolutions.* University of Chicago Press.

Le N.T, Meunier J.-G., Chartrand L. et al. (2016). Nouvelle méthode d'analyse syntactico-sémantique profonde dans la lecture et l'analyse de textes

assistéespar ordinateur (LATAO). In Mayaffre D., et al. (eds.), *Proceedings of the 13th International conference on statistical analysis of textual data.*

Manning C.D. at al. (2009). *Introduction to Information Retrieval*. Online edition. Cambridge, UK: Cambridge University Press.

McKinnon A. (1973). The conquest of fate in Kierkegaard. *CIRPHO*, 1(1): 45-58.

Meunier J.-G. and Forest D. (2005). Classification and categorization in computer assisted reading and analysis of texts. In Cohen H. and Lefebvre C. (eds.), *Handbook of categorization in cognitive science*, pp. 955-978. Elsevier.

Meunier J.-G. and Forest D. (2009). Lecture et analyse conceptuelle assistée par ordinateur: premières expériences. In *Annotation automatique et recherche d'informations*. Hermes.

Mitchell T.M. (1997). *Machine learning*. McGraw-Hill.

Moretti F. (2005). *Graphs, maps, trees. Abstract models for a literary history.* Verso.

Moretti F. (2013). *Distant reading*. Verso.

Pal A.R. and Saha D. (2015). Word sense disambiguation: A survey. *International Journal of Control Theory and Computer Modeling*, vol. 5(3).

Pincemin B. (2007). Concordances et concordanciers: de l'art du bon KWAC. *XVIIe Colloque d'Albi. Langages et signification – Corpus en lettres et sciences sociales: des documents numériques à l'interprétation*, pp. 33-42.

Pulizzotto D. et al. (2016). Recherche de "périsegments" dans un contexte d'analyse conceptuelle assistée par ordinateur: le concept d'"esprit" chez Peirce. *JEP-TALN-RECITAL 2016*, vol. 2, pp. 522-531.

Rastier F. (2011), *La mesure et le grain. Sémantique de corpus*. Champion.

Sainte-Marie M. et al. (2010). Reading Darwin between the lines: a computer-assisted analysis of the concept of evolution in the Origin of species. *10th International conference on statystical analysis of textual data.*

Salton G. (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing*. NJ: Prentice-Hall, Upper Saddle River.

Schmid H. (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees." In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester; UK.

Schmid H. (1995). "Improvements In Part-of-Speech Tagging With an Application To German." In *Proceedings of the ACL SIGDAT-Workshop*, 47–50.

Slingerland E. et al. (2017). The distant reading of religious texts: A "big data" approach to mind-body concepts in early China. *Journal of the American Academy of Religion*: 1-32.