Atti del XLIV Congresso Nazionale Sisfa - Firenze 2024 (2025), **43**_28, pp. 245–252
Doi:10.6093/978-88-6887-352-3-28

Presentation of a 'Mach Corpus' and its preliminary analysis

Enrico Gasco[1]

[1]Zirak S.r.l., Mondovì (Cn), enrico.gasco@zirak.it.

Abstract: In recent years, an increasing amount of digitally available historical texts has become available, and the use of computational tools to explore such masses of sources can be of invaluable help to historians of science. The computational approach has made new tools and models available for historical analysis which have allowed an interpretation of historical texts less linked to the preferences of the scholar. For example, in the history of ideas/concepts, the computational approach has allowed the interpretative models constructed by science historians to be verified in a more precise manner. In this presentation, we want to introduce a corpus of Mach's English-language writings in such a way that it can be used for computational analysis. In particular, the corpus will be annotated for subsequent conceptual analysis. Furthermore, we will try to highlight some characteristics of the corpus as a whole and its initial studies with Machine Learning techniques.

Keywords: Mach, Machine Learning, Computational History

1. Introduction

In recent years, an increasing amount of digitally available historical texts has become available, and the use of computational tools to explore such masses of sources can be of invaluable help to historians of science. Managing the explosion of electronic document archives requires new tools to automatically organize, search, index, and browse large collections.

Using computational tools to explore the history of science opens up exciting possibilities for deepening our understanding of the past. These tools allow historians to manage large data sets, create connections between different disciplines, and interact with historical documents in entirely new ways.

One of the most frequently used computational tools is the one based on Natural Language Processing (NLP) used in studying concepts and their interconnections in large collections of historical texts. For example, Van Wierst et al. (2016) developed a computational approach to analyze the degree of similarity between different books by comparing the number of occurrences of certain key terms. A similar approach was provided by Alfano (2018), who analyzed a set of books by a given author – in his case Nietzsche – taking into account the passages in which certain terms are present. Another interesting way of using NLP was proposed by Betti et al. (2019) and Overton (2013). Both use NLP to identify passages or sequences of words. Still, while the former subsequently uses human experts to classify these passages according to previously defined criteria, the latter applies a series of algorithms on a random selection of articles and then generalises the result to the entire data set. One of the advantages of using computer tools – such as NLP – is that they provide tools to visualize the results (graphs, diagrams, networks) that facilitate subsequent second-level analysis by historians.

The identification of concepts in a given set of texts is facilitated by the machine learning technique of topic modelling; through it, texts are classified based on the topics that represent them. Among the various algorithms that are most used, it is worth mentioning the Latent Dirichlet Allocation (LDA) that for example Hall et al. (2008) used to study the evolution of some ideas in the field of Computational Linguistics in a given period (1978-2006). The topic modelling technique and its use in the social sciences

have been analysed by Pääkkönen and Ylikoski (2021) who highlighted how these techniques allow the discovery of unexpected information in large and diversified corpora, thus improving the transparency of the interpretative process. Starting from 2010, with the work on word-embedding techniques, a 'geometric' approach has been attempted to identify the relationships between concepts in a corpus of texts: an example of this is the project by Bloem et al. (2019) who address the problem of consistency of semantic space by focusing on a dataset that collects Quine's contributions. On a different line – which does not contemplate a historical analysis – is the work by De Sanctis and Rizzi (2023) who use word-embedding algorithms to build conceptual spaces with the task of representing a set of articles downloaded from an online repository. On the use of conceptual spaces in the historical context, the present author also proposed a contribution to a SISFA conference in 2012 (Gasco, 2012), where some passages by Mach and Einstein relating to Mach's Principle were analyzed. One of the difficulties that had been encountered in that project was the small number of texts for an appropriate analysis from a computational/geometric point of view; this article represents a continuation of that project and focuses on the proposal of a Machian Corpus and its initial analysis with machine learning techniques.

2. Mach Corpus

Since the middle of the last century – in computation linguistics and model language – collections of texts have been created – which take the name of corpora – structured in such a way as to be able to be treated with automatic tools. Since the 2000s, numerous corpora have been created in the most varied fields and in languages not strictly related to English (for example, there are corpuses of texts in ancient Greek); in the historical field, some non-strictly standard corpora have been built by recovering texts from online archives (arXiv, Gutenberg project, Jstor) and other more specific ones such as the Royal Society Corpus which includes the articles of the 'Philosophical Transactions of the Royal Society of London' from 1665 to 1920.

In building a corpus – in addition to identifying the texts contained in it – one of the main problems is to establish its representation and which data must be included. In general, the representation is built through meta-data that encapsulates the information that you want to preserve and is implemented through tag-based languages, such as HTML, XML, or JSON so that the corpus can be read by automatic tools. This information can be the most varied; it ranges from the organization of text – subdivision into chapters, paragraphs, sentences – to the syntactic structure of sentences – through post taggers – to end with specific annotations related to the content of some paragraphs. Therefore, the choice of the corpus format is important and we have decided to use the vertical text format (VRT) which represents the input format for the Corpus Workbench (CWB), a set of tools that allow you to efficiently encode and query corpora[1] . As for the information to be saved, we will focus on the text structure, images, formulas, and notes, leaving out the pos tagging information which is the least significant for historical research.

In this article, we want to present a corpus that collects some of Mach's works downloaded from English-language online repositories and formatted in VRT. The works of Mach that we will discuss are the following:

- HCE: *History and root of the principle of the conservation of energy* (1872): traduced by P. E. B. Jourdain (1911)
- PSL: *Popular Scientific Lectures* (1894): traduced by T.J. McCormack (1895)
- PTH: *Principles of the theory of heat – Historically and Critically Elucidated* (1896): traduced by T.J. McCormack (1904)

[1] . There is a command line interface (CQP) and a web-based interface (CQPweb)

- SM: *The science of Mechanics* (1883): traduced by T.J. McCormack (1902)
- AS: *The Analysis of Sensations and the Relation of the Physical to the Psychical* (1905): traduced by C.M. Williams (1914)
- KE: *Knowledge and error – Sketches on the Psychology of Enquiry* (1905): traduced by T.J. McCormack (1926)
- SG: *Space and geometry in the light of physiological, psychological and physical inquiry* (1906): traduced by T.J. McCormack (1907)
- PPO: *The principles of physical optics* (1913): traduced by J.S. Anderson and A.F.A Young (1926)

3. Mach Corpus: construction and first analysis

The construction of a corpus in VRT format requires a rather complex process consisting of several steps that we will list below, avoiding describing in detail the XML structure that is the basis of the VRT.

First of all, all the works must be obtained in text format[2] so that they are processable through Python – the programming language we have chosen. Fortunately, this first point is relatively easy to pursue since the works are downloaded in PDF format, from which it is possible to easily extract the ASCII content[3]. The text is also analyzed by the NLTK library[4] that allows us to divide it into sentences, and tokens and possibly use the library's POS tagging to obtain the fine structure that interests us to build the VRT format. However, there are still some aspects that must be done manually and that require a lot of time: first of all, it is necessary to determine semi-automatically the subdivision into chapters, pages, and paragraphs. Secondly, it is necessary to determine the images, formulas, and notes that are present in considerable quantities in Mach's work. The images are obtained from the text in PDF format through a simple application developed in C#, as well as the formulas, which are typed by hand in Latex format and transformed into images through a simple function: for both types of images, the corresponding file name is indicated in the corpus. Finally, as regards the notes, they must be determined by hand, they are not divided into sentences and determine a paragraph uniquely. All this information is represented in XML format, through tags and attributes, which as mentioned we will not go into detail.

Corpus/Book	N. words	N. sentences
MC (Mach Corpus)	45861	83171
HCE	4503	1160
AS	8824	4323
KE	19343	5011
PSL	9558	4291
PPO	7504	5646
PTH	15717	4083
SM	7253	4852
SG	4584	1538

Tab. 1: General information on Mach corpus

With the documents in VRT format available, it is now possible to give some general information about the Machian corpus. If you apply some simple Python scripts you can obtain the number of unique words in the corpus and the number of sentences present: the data are reported in Tab. 1.

[2] A binary format like MS Word is not usable.

[3] The text is also cleaned from spurious characters.

[4] Downloaded from www.nltk.org

As can be observed, the corpus presents a relatively small number of words compared to the standards, especially for its geometric treatment where a vocabulary with millions of terms is required. Despite this difficulty, we can proceed to determine the most important concepts of the entire corpus and of the individual texts. As we indicated in the introduction, a possible strategy is to use LDA to determine the topics of the corpus, but this algorithm also requires a very large dataset; we, therefore, limit ourselves to using a simpler algorithm such as TFIDF (term frequency-inverse document frequency) which is a function used in information retrieval to measure the importance of a term concerning a document or a collection of documents.

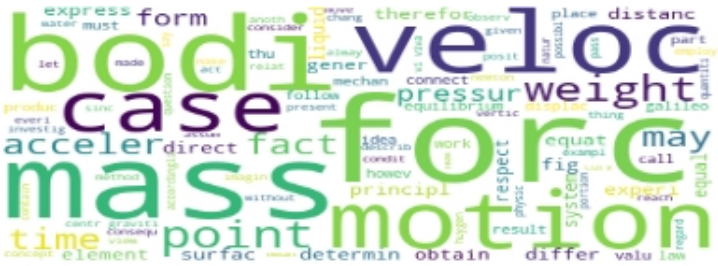

Fig. 1: Words cloud of SM.

It is also necessary to make further assumptions: to focus on significant terms we will not consider the 'stop words'[5] (for example conjunctions), we will use a stemming algorithm[6] and finally, we will use the subdivision into paragraphs used to create the VRT file in order to divide the texts into sub-documents. If we limit ourselves to SM – given the brevity of the intervention – the most significant terms are: 'weight', 'veloc', 'bodi', 'case' 'direct', 'distanc', 'equal', 'equilibrium', 'fact', 'forc', 'form', 'liquid', 'mass', 'may', 'motion', 'point', 'pressur', 'principl', 'time', 'acceler'. As you can see, some words correspond to the main concepts discussed in Science of Mechanics, such as force, mass, and motion. A useful representation of this list is given by the word cloud of the document, where the most significant terms have a larger size, which is shown in the Fig. 1.

We can also ask ourselves how different the texts are from each other; a difference that also indicates the diversity of concepts addressed. To this end – following Degaetano-Ortlieb and Teich (2022) – we can use the relative entropy or Kullback–Leibler Divergence (KLD) which is a widely used method of comparing probability distributions measuring the number of additional bits needed to encode a given dataset A when a (non-optimal) model based on a dataset B is used. The KLD formula is:

$$D\,(A||B) = \sum_i p(item_i|A)\log_2 \frac{p(item_i|A)}{p(item_i|B)} \tag{3.1}$$

where $p(item_i|A)$ is the probability of a linguistic unit in corpus A and $p(item_i|B)$ is the probability of the same unit in corpus B. Note that the formula is not symmetric and therefore, in general, we have that $D(A||B) \neq D(B||A)$; this leads us to choose a particular text – e.g. AS – and then compare it with all the others. The calculation of KLD[7] is shown in the Tab. 2 where the texts most similar to AS are KE and PSL, which are the works with the least technical content.

4. The concept of 'relative' in Mach Corpus

The concept of 'relative' has been studied frequently in the history of ideas and a central role has been given to Machian work. In a previous article (Gasco, 2016) we investigated the concept from a philosophical point of view using the tool of Dynamic Frames. We showed that the concept could be represented with a relation between elements and that this relation was distinguished based on its

[5] Stop words are terms which are filtered out before or after processing of natural language data.

[6] Stemming is the process of reducing the inflected form of a word to its root form, called the "stem".

[7] In the calculation of KLD we used a Jelinek–Mercer smooth function with lambda 0.05 .

BOOK	KLD
AS	0
HCE	1.58
KE	1.10
PLS	0.98
PPO	1.57
PTH	1.46
SM	1.54
SG	1.23

Tab. 2: KLD based on AS

complexity; the simplest form of relation was co-existence, then there was a metric relation (the elements are compared), and finally a functional relation.

Book	N. occur.	Context words
HCE	10	-
AS	116	funtional[4], dependence[4], elements[8], physical[7], mass[8], stand[8], sensation[7], standing[4], body[10], position[8], rotation[4], part[6]
KE	85	fact[4], closely[9], space[8], will[6], have[4]
PSL	61	point[4], closely[4], have[5], motion[4], will[4]
PPO	103	object[12], physical[5], light[14], made[7], depend[4], space[8]
PTH	77	must[5], made[7], same[6], heat[12], equation[4], process[5]
SM	121	things[4], force[13], equal[4], produced[4], have[4], obtain[12], subsist[4], principle[5], time[6], motion[23], acceleration[7], position[6], mass[13], will[6], velocity[25], body[16], absolute[11], universe[4], part[7]
SG	20	-

Tab. 3: 'relative' concept on Mach's books

Following the strategy proposed by Betti (2019), we can build a model of a concept by examining the terms that characterize it and the set of words that fall into the contexts of their use. In this way the concept of 'relative' is determined by a model constituted by the set of occurrences of the following terms: 'relation', 'relative', 'relatives', 'related', 'reference'[8]. As mentioned, we cannot limit ourselves to these single words to frame the concept univocally and since there is no precise definition of 'relative' in Machian work, but its application in different fields and in different examples, we must consider the context of use of these concepts to have a richer representation. Let us therefore examine a context composed of a window of 10 words to the left and right of the target term; in this way, we obtain a series of words that occur with a certain frequency and that show how the concept of 'relative' is characterized by other specifically Machian words that clarify its meaning. In the tab. 3 we report the results obtained for the individual texts, considering a minimum word length of 3 characters[9] and several occurrences greater than 3.

[8] We don't use stemming in this case.

[9] The choice of a maximum word length is to avoid the presence of prepositions that are not significant.

The second column indicates the number of occurrences of the target terms, while the third column specifies the context terms with the relative number of occurrences in square brackets. From Tab. 3 it can be observed that the target terms in some texts are few and the corresponding context words are absent (e.g. HCE), while the texts with a greater number of occurrences are AS, PPO, and SM. It has not been indicated in the table, but the target term that occurs most is 'relation' with a percentage of 55.5%, while the one with the lowest percentage is 'relate' (2.2%).

If we observe the context terms instead, we notice the presence of a few adjectives, some verbs with a precise meaning, and some nouns. Let us try to analyse the context terms in more detail, to have further information on the meaning and use of the concept 'relative'. The adjectives are 'functional', 'physical', 'closely', 'same', and 'absolute'. If we use bi-grams around the target words we notice for example that the adjective 'functional' is always paired with the term 'relation', as can be seen from the following sentence from AS: '. . . *reduce everything to a functional relation of sensational elements*. . . '. Similarly, 'closely' is associated with the term 'related' as evidenced by the sentence extracted from KE: '*the parts of the body are very closely related*'.

More complex is the analysis of the nouns that are present in the relative concept window. To identify the terms that intervene in a relationship we study the bi-grams centered on the word 'relation' to determine some patterns useful for our purpose. The most common bi-grams are listed in the tab. 4 [10]:

Bi-gram	N. occurrence
relation between	46
relation of	73
relation to	33
relation is	13
same relation	11

Tab. 4: bi-gram around 'relation' word

If we consider the bi-gram 'relation between' we note that it establishes a relationship between two entities. We can therefore consider the terms following the bi-gram in a window of a certain size to determine the entities that appear in the relationship most frequently; using a window of 10 words we obtain that the most used terms are 'force', 'space', 'distance' and 'heat'. Among the highlighted terms there is not necessarily a relationship, but we assume they are those that represent the first term of the relationship. For example, if we take 'heat' into consideration and identify the terms that co-occur with it and 'relation between' we obtain that the most significant terms are 'work', 'law', and 'force': we can therefore assume that they represent the second term of the relationship. Using the same procedure, if we consider the term 'space' as the first term of the relationship we obtain that the possible second terms are 'sensation' and 'object'. With this simple methodology, we have determined which terms-concepts intervene in the Machian passages where a relationship between entities is expressed.

At this level it is not possible to establish the type of relationship that exists between the terms that constitute it; let us then try to follow another path to identify whether the types of relationship that he has in mind are indicated in Mach's writings. To this end, let us consider the 'context word surprise' defined as:

$$s_w^d = \log_2 p(w) \tag{4.1}$$

[10] The table does not contain bi-grams which have no particular meaning such as "the relation", "a relation".

where w is the word present in a context of size d, and p(w) is its probability. To interpret the formula, consider that the higher the probability of the word, the lower its surprise value, and similarly, the less frequent the word, the higher its degree of surprise. In the analysis of the contexts of the word 'relation', we will be interested in determining the terms with the lowest surprise value, that is, those that most characterize the relationship under examination. If we consider a context of 3 words around the term 'relation' and determine the degree of surprise for each word that falls within the context, we obtain that the word with the lowest degree of surprise is 'physic'[11], which indicates a physical relationship.

5. Conclusions

In this article, we have presented the strategies for building a Mach Corpus based on documents available online and its preliminary analysis with the tools made available by Machine Learning.

Bibliography

Alfano, M. (2018). "Digital Humanities for History of Philosophy: a Case Study on Nietzsche", in Levenberg, l., Neilson, T. & Rheams, D. (eds.), *Research Methods for the Digital Humanities*. Cham: Palgrave Macmillan, pp. 85-101.

Betti, A. *et al*. (2019). "History of philosophy in ones and zeros", in Curtis, M & Fischer, E. (eds.), *Methodological advances in experimental philosophy*. New York: Bloomsbury Academy, pp. 295-332.

Bloem, J., Fokkens, A. & Herbelot, A. (2019). "Evaluating the consistency of word embeddings from small data", in Mitkov R. & Angelova, G. (eds.), *Recent Advances in Natural Language Processing*, International Conference on Recent Advances in Natural Language Processing, RANLP 2019, Varna, Bulgaria, September 2-4, 2019, pp. 132-141. Available at: acl-bg.org (Accessed on 19 November 2024).

De Santis, E. & Rizzi, A. (2023). "Prototype Theory Meets Word Embedding: a Novel Approach for Text Categorization via Granular Computing", *Cognitive Computation*, 15(3), pp. 976-997.

Degaetano-Ortlieb, S. & Teich, E. (2022). "Toward an optimal code for communication: the case of scientific English", *Corpus Linguistics and Linguistic Theory*, 18(1), pp. 175-207.

Gasco, E. (2012). "Semantic Spaces and History of Physics: a case study", in *Abstracts of XXXII Congresso SISFA 2012*.

Gasco, E. (2016). "The concept of relativity in Mach". Available at: pitt.edu (Accessed on 18 November 2024).

Hall, D. & Jurafsky, D. & Manning, C.D. (2008). "Studying the history of ideas using topic models", in Lapata, M & Ng, H.T. (eds.), *Proceedings of the 2008 conference on empirical methods in natural language processing*, Honolulu, Hawaii, pp. 363-371.

Pääkkönen, J., & Ylikoski, P. (2021). "Humanistic interpretation and machine learning", *Synthese*, 199(1), pp. 1461-1497.

Overton, J.A. (2013). " 'Explain' in scientific discourse", *Synthese*, 190, pp. 1383-1405.

Van Wierst, P., *et al*. (2016). "Phil@Scale: Computational Methods Within Philosophy", in Wieneke, L. *et al.* (eds.), *Digital Humanities Luxembourg*, 3rd Conference on Digital Humanities, Luxembourg, December 5-6. Available at: ceur-ws.org (Accessed on 19 November 2024).

[11] In this procedure we used the Porter algorithm for stemming.