Received: 30 September 2022 | Accepted: 18 May 2023

DOI: 10.1002/fhu2.6

ORIGINAL PAPER

future humanities WILEY

Machine learning for the history of ideas

Simon Brausch[1] | **Gerd Graßhoff**[2]

[1]Max Planck Institute for the History of Science, Berlin, Germany

[2]Max Planck Institute for the History of Science, Berlin Institute for the Foundations of Learning and Data (BIFOLD), Humboldt Universität zu Berlin, Berlin, Germany

Correspondence

Simon Brausch, Max Planck Institute for the History of Science, Boltzmannstr. 22, Berlin 14195, Germany.
Email: sbrausch@mpiwg-berlin.mpg.de

Abstract

The information technological progress that has been achieved over the last decades has also given the humanities the opportunity to expand their methodological toolbox. This paper explores how recent advancements in natural language processing may be used for research in the history of ideas so as to overcome traditional scholarship's inevitably selective approach to historical sources. By employing two machine learning techniques whose potential for the analysis of conceptual continuities and innovations has never been considered before, we aim to determine the extent to which they can enhance conventional research methods. It will amount to a critical evaluation of how the advantages of computational in-breadth analysis could be combined with the merits of traditional in-depth analysis in a philosophically fruitful way. After a brief technical description, the approach will be applied to an example: the conceptual (dis)continuity between medieval and early modern philosophy. All the challenges encountered during development and application will be carefully evaluated. We will then be able to assess whether these tools and techniques present promising extensions to the methodological toolbox of traditional scholarship, or whether they do not yet have the potential for a task as complex as the analysis of philosophical literature. The present investigation can thus be seen as an experiment on how far one can go with current machine-learning techniques in this area of research. In doing so, it provides important insights and guidance for future advances in the field.

KEYWORDS

computational philosophy, digital humanities, history of ideas, machine learning, natural language processing

1 | INTRODUCTION

The information technological progress that has been achieved over the last decades has also given the humanities the opportunity to expand their methodological toolbox. In philosophy, this has led to the emergence of what is usually referred to as computational philosophy. The present investigation aims at contributing to this field by experimenting with a computational approach to the evaluation of conceptual continuity and innovation in the history of philosophical ideas. Investigating the history of philosophy with the help of computational means is primarily motivated by the fact that a human scholar alone cannot possibly deal with the vast amount of historical material that may be relevant for any given research question. Before the investigation even begins, one must narrow down the number of authors and writings that are to be considered at all to a humanely manageable size. As a consequence, the traditional in-depth analysis of terminological and conceptual shifts used to track (dis) continuities in the history of ideas has always been restricted in its breadth.[1] Hence, it only seems natural to take advantage of the ever-growing processing power of modern computers to overcome traditional scholarship's inevitably selective approach to historical sources. But being in possession of the technological means does of course not yet provide a straightforward answer to the question of *how* the available processing power should be applied to attain philosophically meaningful results. So how exactly might the use of computational means enhance, extend or even replace traditional research methods?

For the history of ideas, one of the most promising areas of computer science is a subfield called *natural language processing* (NLP).[2] The approaches made so far to apply recent advancements in NLP to research on historical questions in philosophy can roughly be divided into two groups.[3] Van Wierst et al. (2016), for example, develop a computational approach for analysing the degree of similarity between different books by comparing the number of occurrences of certain previously specified key terms. A similar approach was provided by Alfano (2018), in which a variety of books from one author, in his case Nietzsche, are analysed with respect to the number of passages per book that contain certain terms. Yet another interesting way to use the NLP-toolbox in this area of research was proposed by Betti et al. (2019) on the one hand and by Overton (2013) on the other. Both apply NLP to identify passages or word sequences in which a certain term occurs. But while the former subsequently make use of human experts to classify those passages according to previously defined criteria, the latter conducts a close-reading session of a randomly sampled selection of the research papers at stake and then generalises the resulting assessment to the whole set. All the other NLP-based approaches to the history of ideas that we are aware of[4] make use of the same or at least very similar kinds of strategies: they employ frequency and concordance analyses with respect to the occurrence of a particular set of key terms and corresponding patterns to either (i) visualise these results in form of graphs, schemes or networks, or (ii) facilitate a subsequent evaluation of the resulting textual data by means of traditional research methods.

The computational approach to the history of philosophical ideas developed in here may be placed into the second group. It was designed for conducting research on questions of

[1]Note that there is a difference between shifts on the terminological and shifts on the conceptual side: one may, for example, introduce a new term for expressing an old idea (terminological shift without conceptual shift) or, inversely, use the same term for a new idea (conceptual shift without terminological shift). While the former may be a sign for conceptual continuity, the latter may indicate conceptual discontinuity, that is, innovation.

[2]As the name already suggests, NLP is concerned with the development of computational techniques and methods for the automatic processing of natural language.

[3]In fact, some of the papers that we discuss in the following are concerned with developing computational approaches to the history of *scientific* ideas. However, as the methods of investigation for the historical development of scientific and philosophical ideas are very similar, computational approaches to the former seem equally relevant for the latter and vice versa. For the sake of brevity, we will refer to both as computational approaches to the history of (philosophical) ideas.

[4]These are: Aiello and Simeone (2019), Gibson and Ermus (2019), Laubichler et al. (2013), and Andow (2015).

conceptual continuity and innovation between different philosophical eras or traditions. Conceptual continuity between traditions is operationalised as conceptually continuous ways in which philosophers belonging to these traditions characterise certain philosophical key concepts.[5] So if the ways two philosophers of different traditions characterise the concept in question are conceptually more continuous to each other than to a third author belonging to either tradition, then we may assume some degree of conceptual continuity between these two traditions.[6] Analysing how certain concepts were used at different points in the history of philosophy may thus lead to the discovery of new connections which force us into giving up long established distinctions or prompt us to draw new lines within traditions or eras previously conceived as one.

The present investigation can be seen as an experiment on how far one can go with current advances in NLP in this area of research. By employing two machine learning techniques whose potential for the analysis of philosophical literature has never been explored before, our goal is to determine whether they may help us to overcome the limitations of traditional scholarship. It will amount to a critical evaluation of how computational in-breadth analysis could be combined with traditional in-depth analysis in a philosophically fruitful way.

The main part of this paper is divided into three sections. In the first section, we will provide a brief technical description of the method.[7] The basic idea is to employ the open-source NLP-library Spacy[8] to conduct a large-scale computational analysis of *all* the writings that may be relevant for a given research question and to identify exactly those text passages in which the authors provide their own view on the concept at stake. To do so, the approach takes advantage of modern machine learning techniques for the development of two specialised models for text classification and named entity recognition that (i) exclude sentences in which the authors are not expressing their own but someone else's point of view and (ii) to keep only those parts of the remaining sentences in which the authors characterise the concept by means of a definitional attribution. The resulting sets of definitional attributions are then transformed into a synoptic presentation which, as we will show, has the potential to drastically facilitate the philosophical analysis of terminological and conceptual shifts through which continuities and innovations in the history of philosophy can be tracked down.

In the next section, the approach will be applied to an example: the conceptual continuity between medieval and early modern philosophy. The idea here is not only to illustrate how the method can be put into practice, but also to verify if the approach works. In the example, the writings of Descartes, Suárez, and Ockham are going to be analysed with respect to the way the philosophical concept of soul is characterised throughout their respective work. We will compare the insights gained through our computational approach with the results of Perler (2013) traditional analysis of these three authors with regard the same research question.

In the last section, we will examine how exactly the approach developed in here complements traditional research methods and conduct an extensive critical evaluation in which all the challenges encountered during development and application will be addressed. We will then be able to assess whether the tools and techniques that we made use of present promising extensions to the methodological toolbox of traditional scholarship or whether they

[5]Conceptually discontinuous characterisation would thus amount to the affirmation of conceptual innovation. In stating this operationalisation, we follow Dominik Perler's approach to conceptual continuity and innovation in the history of philosophical ideas which will be introduced in more detail in Section 3 below.

[6]As soon as one compares the writings of more than two authors, we assume conceptual continuity and innovation to come in (not explicitly quantifiable) degrees: even though there may be some conceptual discontinuities (i.e., some innovation) between the ways two authors characterise a certain concept, they may nonetheless be conceptually *more* continuous to each other than to the characterisation of a third author.

[7]In the following, we will use the terms 'method' and 'approach' synonymously.

[8]Git-repository: https://github.com/explosion/spaCy [Accessed 13th September 2022]; for the present paper, Spacy version 2.3.2 was used.

do not yet have the potential for a task as complex as the analysis of philosophical literature. We will draw on these insights to formulate recommendations and directives for future research.

2 | METHODS

In this section, we will provide a brief technical description of the computational approach proposed in this paper. It was developed in the Python 3 programming language (version 3.8.6) and carried out in a Jupyter Notebook environment.[9] It is essentially composed of four consecutive workflows that correspond to four different Jupyter Notebooks, hereafter just referred to as notebooks. Every notebook comes with an explanatory section at the top as well as indications for all adjustments that are required when applying the templates to a specific research project.[10]

2.1 | OCR-notebook

To carry out a computational analysis, all the relevant writings of the authors in question must first be obtained in digital form.[11] After the data has been collected, the task of the first notebook is to produce a plain-text version of those writings which are only available in PDF. The notebook starts by preprocessing the text: first, all the irrelevant material—such as the front page, the table of contents, and the index—is cut out with the help of the Python library PyPDF2.[12] The remaining pages are then converted into jpegs with the Python library pdf2image.[13] After these two preliminary steps, the Python API pytesseract[14] is used to run tesseract[15]—an open-source OCR-engine—on all the images one by one. To keep only the main body of text in the final plain-text version of the respective writing, three additional criteria for layout analysis were developed for detecting extraneous material such as headers, footnotes, and page numbers. As some books are structured in *numbered* paragraphs while others are not, there are two versions of the OCR-notebook which only differ with respect to the last of the three criteria:[16] the first version of the OCR-notebook makes use of what we call the *footnote criterion* and is designed for books that are *not* structured in numbered paragraphs. The second version is designed for books that *are* structured in such a way and makes use of what we call the *font size criterion*.[17] As a result, the three criteria taken together are capable of eliminating almost all the extraneous material and thereby enable the OCR-engine to generate a more or less flawless plain-text version of the respective writings. Note that the criteria for layout analysis were not merely developed for aesthetic reasons. Obtaining reasonably clean plain-text versions of the writings' main body of text is probably the most essential precondition for the success of the entire project.

[9]Git-repository: https://github.com/jupyter/notebook [Accessed 13th September 2022].
[10]The notebook templates are included in the supporting material for this paper.
[11]In a best-case scenario, all the writings of a specific author are available in digital form and they can thus *all* be taken into account for the computational analysis. But since at present, this is often not yet the case, it is important to get hold of at least those writings which are known to be particularly relevant with respect to the research question.
[12]Git-repository: https://github.com/mstamy2/PyPDF2 [Accessed 13th September 2022].
[13]Git-repository: https://github.com/Belval/pdf2image [Accessed 13th September 2022].
[14]Git-repository: https://github.com/madmaze/pytesseract [Accessed 13th September 2022].
[15]Git-repository: https://github.com/tesseract-ocr/tesseract [Accessed 13th September 2022].
[16]For more details on the three criteria for layout analysis, we refer the reader the notebook templates.
[17]After testing the performance of the two different versions of the third criterion on several books, it was found that the *footnote criterion* is slightly more reliable in correctly detecting footnotes than the *font size criterion*. Hence, the latter version should only be used for books that are structured in numbered paragraphs.

In addition to the three criteria for layout analysis, a further mechanism was developed for removing sentences that contain low-confidence words. Tesseract assigns a value between 0 and 100 to every word, indicating the degree of confidence with which the word was recognised correctly. While joining the recognised text as string, the mechanism replaces all low confidence words, that is words with a confidence value inferior to a previously determined threshold, with an easily identifiable 'remove word' and subsequently removes sentences containing this word.[18]

2.2 | NLP-preparation-notebook

The next notebook takes the writings in plain-text format as input and uses the NLP-toolbox provided by Spacy to identify and extract every sentence in which the philosophical key term—whose varying characterisations among the different authors is at stake—occurs. To do so, Spacy's matcher-function is used to define a match pattern for the philosophical key term in question.[19] After eliminating potential duplicates, the list of sentences containing a match is written to a jsonl-file and the author's name and the title of the source are added as metadata. As a result, the final output of this notebook is a jsonl-file containing all the sentences of one text in which the term at stake occurs. Every line of the jsonl-file is composed of two name-value pairs: (i) a sentence and (ii) the metadata indicating its source.

2.3 | NLP-analysis-notebook

The aim of this notebook is to single out all instances in which the respective author characterises the concept at stake by means of a definitional attribution. To do so, two NLP-models were developed specifically for this purpose. They constitute the key components of the present approach. To train the two models, the annotation tool Prodigy[20] was run on the data generated by the nlp-preparation-notebook in the course of the example application that will be presented in Section 3 below. Hence, during development the concept whose varying characterisations among different authors was at stake was the concept of *soul*.

The notebook is essentially composed of three consecutive operations. After having converted the jsonl-file from the previous notebook into a Pandas DataFrame,[21] hereafter just referred to as dataframe, a specialised NLP-model for text classification is first being used to eliminate all sentences in which the respective author is not expressing his or her *own* but *someone else's* point of view on the philosophical concept at stake. In other words, for every given sentence in which the term occurs, the model's task is to predict if it presents an instance of what we call a *third-party ascription*, that is, a characterisation of the concept that is ascribed to a third-party author or source, by assigning a value between 0 and 1. For this purpose, the text classification label 'TP_ASCRIP' (third-party ascription) was created. Prodigy's manual annotation recipe *textcat.manual* was then being used in the classification interface to annotate the data: for every sentence it was decided if the label applies or not. Whether the model is capable of learning from the annotated data and, by implication, whether the model is going to

[18]For instance, in the example described in the next section, the threshold was set to 40 as observation showed that words with a confidence value of more than 40 were usually recognised correctly.
[19]The match pattern can also include terms that are known or suspected to be used synonymously with the term in question by the respective authors.
[20]Prodigy is a commercial annotation tool for machine teaching created by ExplosionAI GmbH; official website: https://prodi.gy/ [Accessed 13th September 2022].
[21]Git-repository: https://github.com/pandas-dev/pandas [Accessed 13th September 2022].

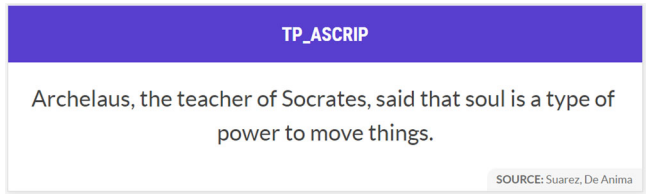

FIGURE 1 Example of a third-party ascription.

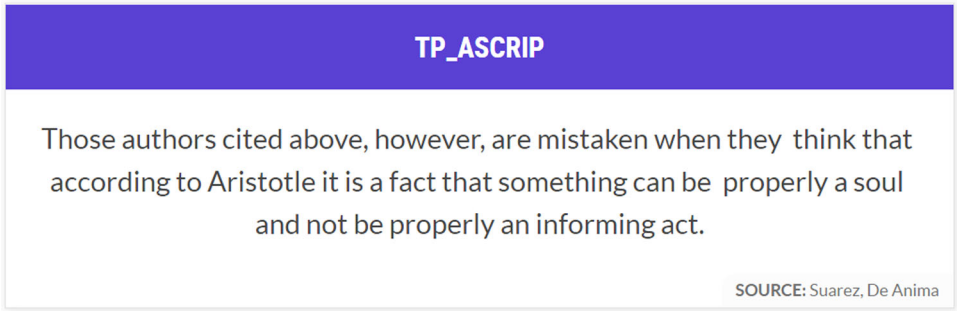

FIGURE 2 Example of a second-order third-party ascription.

be useful at all for the present purpose, essentially depends on the degree of internal consistency with which the annotation decisions were made.

First and foremost, the label was used for (i) unequivocal cases of third-party ascriptions such as the one shown in Figure 1, which are easily identifiable by formulations of the type 'author X asserts that …', 'according to author X, …,' or ' … as author X said'. In addition, the label was applied to sentences containing (ii) citations in quotation marks, (iii) references to indeterminate others (e.g., 'some', 'most'), (iv) references to the Bible, common faith or knowledge, (v) inferences to be rejected (e.g., 'the following proof is fallacious: …'), and (vi) what we call second-order third-party ascriptions: instances in which the author expresses a third-party author's view on a fourth-party author's characterisation of the concept. Figure 2 below shows an example.

After 500 sentences were annotated according to this annotation policy, Prodigy's train recipe was used in conjunction with Spacy's large English model 'en core -web lg' to train a specialised model for third-party ascriptions. While 350 of the 500 annotated sentences were used for training, the remaining 150 sentences (30%) were held back for evaluation—for which the model reached a ROC-AUC score of 0.977. When applied to the dataframe in the nlp-analysis-notebook, the model now provides a prediction value for third-party ascriptions for every given sentence (Table 1).

Second, a specialised NLP-model for named entity recognition (NER) is run on every row of the dataframe. Its task consists in identifying instances in which the term at stake is characterised by means of a definitional attribution.[22] For every given sentence in which the term occurs, the model is supposed to predict if a definitional attribution appears *in* the sentence—and if so, where exactly it appears.

[22]By 'definitional attribution' we denote instances in which a characterisation is attributed to the term in an affirmative way. For more details, see the annotation policy discussed below.

TABLE 1 Example of a tp_ascrip-dataframe, nlp-analysis-notebook, step 2.2, Suárez's *De Anima*.

	Text	Meta	tp_ascrip
…	…	…	…
4	But the soul does indeed suppose a complete substance.	{'source': 'Suarez, De Anima'}	0.002623
…	…	…	…
13	Albert in his commentary on this passage, chapter 6, says that to live is the act of a soul in that which is animated.	{'source': 'Suarez, De Anima'}	0.994970
…	…	…	…

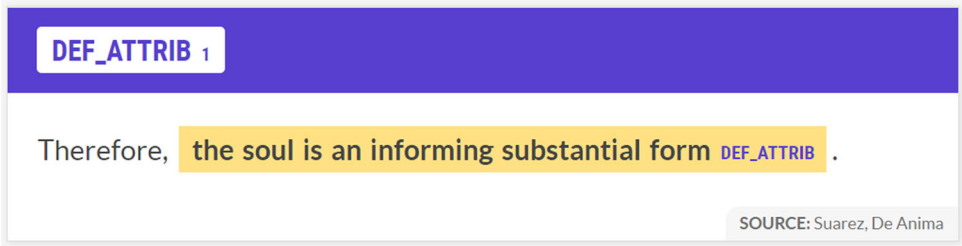

FIGURE 3 Example of a definitional attribution.

For this purpose, the NER-label 'DEF_ATTRIB' (definitional attribution) was created. Prodigy's manual annotation recipe *ner.manual* was then used in conjunction with Spacy's large vector model[23] to annotate the data: now, for every sentence it was decided if the label applies to a subsection of the sentence. If so, the relevant subsection, that is, the entity span expressing a definitional attribution, was tagged. As in the case of the text classification model, a consistent annotation policy is key to success.

In addition to (i) rather obvious case such as the one presented in Figure 3, the label was also used for (ii) definitional attributions of all *subclasses* of the term at stake (e.g., the *vegetative* soul), but *not* for instance in which the term appears in form of a genitive construction in which it is not the term itself but some of its qualities or parts that is being characterised (e.g., the *grade of* the soul). The label was further applied to (iii) instances of a definitional attribution *by* negation ('the soul is not simple'), but not to *negated* definitional attributions ('I do not intend to teach that all souls are …'). Cases of (iv) deductive inferences were also labelled as definitional attributions, whereas conditional clauses and characterisations by means of certain modal verbs ('ought', 'should') were not.[24]

After 655 sentences were annotated according to this annotation policy, Prodigy's train recipe was used in conjunction with Spacy's large vector model 'en vectors web lg' to train a specialised model for definitional attributions. On the 30% of annotated data that was held back for evaluation, the model reached an F-score of 72.34—with precision at 75 and recall at 69.86. In the nlp-analysis-notebook, applying the NER-model now adds a new column to the previous dataframe (Table 2).

Now that a prediction value for third-party ascription has been assigned to every sentence and cases of definitional attributions have been identified, the resulting dataframe is sorted by

[23]The model is available on the official website of Spacy, URL https://spacy.io/models/en#en vectors web lg [Accessed 13th September 2022].
[24]The reason for this is that conditional clauses as well as ought- and should-statements are not (yet) affirmative of the concept at stake.

TABLE 2 Example of an unsorted dataframe, nlp-analysis-notebook, step 3.2, Suárez's *De Anima*.

	Text	Meta	tp_ascrip	def_attrib
…	…	…	…	…
4	But the soul does indeed suppose a complete substance.	{'source': 'Suarez, De Anima'}	0.002623	[the soul does indeed suppose a complete substance]
…	…	…	…	…
13	Albert in his commentary on this passage, chapter 6, says that to live is the act of a soul in that which is animated.	{'source': 'Suarez, De Anima'}	0.994970	[]
…	…	…	…	…

the prediction value to find an appropriate threshold for the reliable elimination of third-party ascriptions. As we are only interested in definitional attributions that express the author's own point of view, the third operation then consists in creating a new dataframe that keeps only those rows in which a definitional attribution occurs and that have *not* been identified as third-party ascriptions according to the previously determined threshold.

Since the NER-model's predictions are not always accurate,[25] the nlp-analysis-notebook also comes with the option to eliminate false positives, that is, rows in which the model has incorrectly identified an instance of a definitional attribution. The final results are then written to a jsonl-file which consists of three name-value pairs per line: (i) the definitional attribution that was found, (ii) the relevant metadata, that is, the respective author and writing, and (iii) the corresponding sentence in which the definitional attribution was found.

2.4 | NLP-comparison-notebook

The last notebook combines the results of all the previous steps by transforming the resulting sets of definitional attributions into a synoptic presentation. To do so, the values for 'definitional attribution' of the jsonl-files that were obtained from the application of the previous notebook to the authors in question are simply converted into a new dataframe. The dataframe is thus composed of multiple columns—one per author—which contain all the cases of definitional attributions identified in the course of the computational analysis. As will be shown in the next section, this presentation may then serve as a basis for the analysis of conceptual (dis)continuities between the ways the authors under investigation characterise the concept at stake.

3 | RESULTS

Now to find out whether the approach works, it was applied to an already well-explored research question in the history of philosophical ideas: the conceptual (dis)continuity between medieval and early modern philosophy. By comparing the results obtained through our computational approach with the results obtained through traditional research methods, it will then be possible to assess its performance. Hence, the purpose of the following application is

[25]The implications of the model's (in)accuracy for future research will be discussed in Section 4 below.

not necessarily to discover anything new but to verify if the insights achieved by a recognised expert for medieval and early modern philosophy who is well acquainted with all the relevant historical sources can likewise be achieved by solely relying on the output produced by the computational approach developed in this paper. So instead of going beyond traditional scholarship in terms of the number of sources included, the experiment presented in this section enables us to identify the potential as well as the limits of using specialised NLP-models for research in the history of ideas. Rather than obfuscating the need for human intervention before, during, and after the use of computational methods, we will document the execution of each step and the challenges encountered along the way in great detail. In this way, we hope to make the subsequent evaluation of strengths and weaknesses in Section 4 as transparent as possible. The detailed documentation will also make it easier for future research to reproduce the results, advance the workflow, and to expand it for the exploration of other research questions in the history of ideas.

3.1 | Research question

Descartes is commonly conceived as the founding father of modern philosophy, a new era in the history of philosophy which put an end to the scholastic tradition of medieval thinkers. This view, however, has recently been challenged by Dominik Perler who claims that in spite of Descartes' attempt to overcome scholastic theories of soul such as Francisco Suárez's account, he nevertheless remains in the medieval paradigm. This is because, according to Perler, Descartes' seemingly new approach to the theory of the soul can in fact be traced back to the work of William of Ockham, that is, yet another medieval scholar (Perler, 2013, p. 34). As a consequence, this assumption gives rise to a more continuous perspective on the history of philosophy by devaluating the role usually attributed to 'great minds' such as Descartes. In other words, it opens up the possibility to rethink the history of philosophy by considering Descartes only as one—nonetheless highly important—contributor to the progressive development from medieval to early modern philosophy. According to this view, the line should not be drawn between medieval and early modern debates, but rather between various competing conceptual approaches originating from the medieval era which were still the subject of philosophical discussion in the early modern period (Perler, 2013, p. 37f.).

The task of the following experiment is to re-evaluate the role of Descartes' work in the transition from scholasticism to modern philosophy with the help of the computational approach introduced above. The idea is thus to take into account all the relevant writings of Ockham, Suárez, and Descartes and to process them according to the different steps outlined above. The resulting data, that is, the sets of definitional attributions of which only small extracts are shown in this paper, are included in the supporting material.

3.2 | Data collection and OCR-notebook

Having formulated the research question and identified the corresponding research object, that is, the authors whose writings are to be analysed, the first step consisted in collecting the data. As Prodigy did not yet provide support for the Latin language when the two specialised NLP-models were developed, it was decided that for the present purpose of illustration and experimentation, it would be sufficient to work with the English translations of the relevant writings.

To determine which writings of the three authors are relevant to the research question, the paper in which Perler puts forward his arguments served as guidance. For a source to be included in computational analysis, it must be available in digital form and in a natural

language for which a suitable NLP-library has already been developed. For the writings mentioned by Perler that met these two criteria at the time the experiment was conducted, the following English translations were used: Suárez, (2011) *Selections from De Anima*, Ockham, (1991) *Quodlibetal Questions*, and Descartes' (1985) *Meditations on First Philosophy*, *Principles of Philosophy*, *The Passions of the Soul*, and *The Search for Truth*. To create a plain-text version of these writings, the OCR-notebook with the *font size criterion* was applied to Descartes' *The Passions of the Soul* and to his *Principles of Philosophy* as they are structured in numbered paragraphs. For the remaining writings, the OCR-notebook with the *footnote criterion* was made use of.

After a first run, both the output information and the output visualisation were examined to verify if the processing was on the right track. In addition, the resulting txt-files were checked for systematic OCR-errors. If necessary, the insights gained from the first run were then used to adapt some of the parameters to the book-specific requirements.[26] The OCR-notebooks were then run a second time on the writings. As a result of the previous fine-tuning, the notebooks were now able to produce almost flawless plain-text versions.

3.3 | Preparation, analysis, and comparison

As described in Section 2 above, the next step consisted in applying the nlp-preparation-notebook to the txt-versions of the respective writings to identify all the sentences in which the term soul occurs. Table 3 below provides a summary of the first intermediate results: it shows the total number of matches found per book and per author, that is, how many times the term soul appears.

At the end of the every nlp-preparation-notebook, the result of this preparatory analysis, that is, all the sentences containing a match, was written to a jsonl-file and the respective metadata was added. These files were then merged into one file per author and subsequently used as input for the respective nlp-analysis-notebooks.

After applying the two specialised models for text classification and named entity recognition to the respective data sets in the nlp-analysis-notebooks, a threshold for the prediction value of third-party ascriptions had to be determined. Although in the training experiment, the text classification model reached a ROC-AUC score of 0.977—that is, a near-perfect accuracy in predicting cases of third-party ascriptions—, the appropriate threshold for eliminating actual cases of third-party ascriptions varied significantly among the three authors.

In the case of Suárez, the vast majority (121) of the sentences in which a third-party ascription occurs (149) did in fact receive a prediction value of 0.9 and above, many (83) even above 0.99, whereas sentences in which the author expresses his own point of view (227) received a value below 0.05, most (211) even below 0.01. However, there were some rare cases (7) of third-party ascriptions to which the model assigned a value between 0.05 and 0.1. For this reason, the threshold for elimination was put at 0.05 (Table 4).

For Ockham, however, the situation was quite different. While most (70) of the negative cases (90), that is, where the author expresses his own view, did again receive a value of 0.01 and below and all cases (8) of third-party ascription a value of 0.9 and above, there were six negative cases to which a value over 0.1 was assigned to—two of them even above 0.8.

As can be observed in Table 5, the outliers can probably be explained by the fact that in these cases, Ockham is talking about the qualities of Christ's soul as well as the role God plays in the capacity of sensation in human souls. Hence, the text classification model seems to have mistaken Ockham's view *on* God and Christ for a third-party ascription *to* God and Christ. Further

[26]In the case of Ockham's *Quodlibetal Questions*, for example, the threshold for removing blocks with the first layout criterion was lowered to 10 words per block as the initial threshold of 16 words turned out to be too invasive. For most of Descartes writings, the item.replace function was used to correct the systematic OCR-error of mistaking the first-person pronoun 'I' for a squared bracket.

TABLE 3 Number of total matches per book and per author.

Author	Writing	Matches/book	Matches/author
Descartes	Meditations on first philosophy	5	275
	Principles of philosophy	20	
	The search for truth	8	
	The passions of the soul	242	
Suárez	De Anima	461	461
Ockham	Quodlibetal question	135	135

TABLE 4 Suárez, extract of sorted dataframe at threshold, nlp-analysis-notebook, step 3.3.

	Text	Meta	tp_ascrip	def_attrib
…	…	…	…	…
341	Moreover, if the soul dies, how can everlasting punishment in Hell be proclaimed?	{'source': 'Suarez, De Anima'}	0.042631	[]
347	Thus someone could say that our soul will last forever because God will preserve it, not because it is intrinsically immortal.	{'source': 'Suarez, De Anima'}	0.050980	[our soul will last forever, it is intrinsically immortal]
31	They thought that God is the soul of the world.	{'source': 'Suarez, De Anima'}	0.058383	[]
…	…	…	…	…

confusion was caused by Ockham's particular style of regularly including auto-references to his previous claims. As a consequence, the threshold for elimination was put at 0.85.

In the case of Descartes, the model's predictions were misled by similar reasons. Again, most (205) of the negative cases (218) correctly received a value of 0.01 and below. However, while the only three cases of third-party ascriptions erroneously received a value below 0.01, there were also eight negative cases to which a value above 0.7 was assigned to—five of them even above 0.9 (Table 6).

A possible explanation may be that in these sentences, Descartes characterises the passions of the soul *in general* by referring to *one particular* but indefinite soul as a means of exemplification. In doing so, the author makes use of the third-person pronoun 'he'—which the text classification model seems to have mistaken for an indication of a third-party ascription. As a consequence, the threshold was put at 1.0 which effectively deactivates this elimination criterion. The three cases of third-party ascriptions which received a value below 0.01 were removed manually in the next step, that is, together with the false positive cases of definitional attributions.[27] What this examination of the model's performance clearly shows is that, due to

[27]Although the clue for the presence of a third-party ascription in these cases was a bit tricky ('as for the opinion of those who think …', 'it has been thought that …', and 'it has been believed that …'), it is nonetheless surprising that the model assigned such a low prediction value to these sentences.

TABLE 5 Ockham, extract of outliers in sorted dataframe, nlp-analysis-notebook, step 3.3.

	Text	Meta	tp_ascrip	def_attrib
…	…	…	…	…
49	For if (i) a sensation exists subjectively in the intellective soul and (ii) God is able to conserve every accident in its subject in the absence of everything else, then it follows that he is able to conserve a sensation in a separated soul—which is absurd.	{'source': 'Ockham, Quodlibeta'}	0.827078	[]
87	But Christ's soul is not able to move its body in heaven nonorganically.	{'source': 'Ockham, Quodlibeta'}	0.837630	[Christ's soul is not able to move its body in heaven nonorganically]
…	…	…	…	…

TABLE 6 Descartes, extract of upper end in sorted dataframe, nlp-analysis-notebook, step 3.3.

	Text	Meta	tp_ascrip	def_attrib
…	…	…	…	…
196	But because the defects in the soul trouble a person only in so far as he becomes aware of them, you have an advantage over us in that, unlike us, you do not notice all the many things which you lack.	{'source': 'Descartes, Truth'}	0.710414	[]
…	…	…	…	…
162	Nevertheless he feels at the same time a secret joy in his innermost soul, and the emotion of this joy has such power that the concomitant sadness and tears can do nothing to diminish its force.	{'source': 'Descartes, Passions'}	0.980563	[]
…	…	…	…	…

the complexity of the philosophical literature in general and the diversity of style in particular, there is no one-fit-all solution.[28] Training the text classification model on more data might decrease the number of outliers, but without considerable advances in machine learning, there will always remain the need to examine the results and adapt the threshold accordingly.

After filtering out the positive cases of third-party ascriptions according to the previously determined thresholds, the remaining sentences in which a definitional attribution was found were gathered in a dataframe (Table 7).

Finally, the dataframe was examined for false positives, that is, for rows in which the model for named entity recognition incorrectly identified an instance of a definitional

[28]This observation already points to a potential limitation of using text classification models in this area of research whose implications will be further discussed in Section 4.

TABLE 7 Ockham, extract of dataframe with false positives, nlp-analysis-notebook, step 3.4.

	def_attrib	Meta	Text	tp_ascrip
0	The intellective soul is the form of the body	{'source': 'Ockham, Quodlibeta'}	Can it be demonstrated that the intellective soul is the form of the body?	0.002214
1	The intellective soul is an incorruptible form	{'source': 'Ockham, Quodlibeta'}	For the opposite: The intellective soul is an incorruptible form.	0.002130
…	…	…	…	…

attribution.[29] Potential candidates for false positives were cases of definitional attributions that were clearly at odds with the annotation policy used in creating the training data for the model—for example, line 0 in Table 7 above which clearly is a non-affirmative question—, as well as entities with very inaccurate end points—especially very long phrases that had been incorrectly detected as one entity. It turned out that it was particularly difficult for the model to identify negated definitional attributions as well as if- and ought-statements as true negatives. In the case of Suárez, 24 out of 116 cases of definitional attributions turned out to be false positives, for Ockham 9 out of 50, and for Descartes 6 out of 29.[30] The three resulting dataframes containing *all* the definitional attributions *without* false positives and third-party ascriptions were then written to a jsonl-file and used as input for the final notebook. Although the number of false positives might be reduced by further improving the annotation policy used to create the training data, it seems unlikely that the verification for false positives will ever become completely redundant.

With the three sets of definitional attributions at hand, the nlp-comparison-notebook was now applied to generate the synoptic presentation. Table 8 below shows a small extract.[31]

3.4 | Interpretation

The final dataframe provides a synoptic presentation of *all* the definitional attributions identified in *all* the writings that were taken into account.[32] It will now be used as a basis for evaluating the conceptual (dis)continuities between the ways Suárez, Ockham, and Descartes conceive of the concept of soul. If Descartes' and Ockham's accounts, that is, the accounts of a medieval and an early modern philosopher, are found to be conceptually continuous, then this may serve as an indication for a certain degree of conceptual continuity between medieval and early modern philosophy. If, in addition, the accounts of Descartes and Ockham turn out to be conceptually *more* continuous than the accounts of Ockham and Suárez, then this may indicate that the line should not be drawn between medieval and early modern debates, but rather

[29]Note that here, we are referring to false positive cases of *definitional attributions*, whereas just before, we distinguished between positive and negative cases of *third-party ascriptions*. Also note that while looking for false positives, some last minor OCR-errors—for example, rare cases of sentences sticking together because of missing punctuation—were corrected manually and the two models were applied again.

[30]Thus, in total, 39 out of 195 cases of definitional attributions were identified as false positives, corresponding to an average false positive rate of 20%.

[31]The entire dataframe is included in the supporting material for this paper.

[32]There may of course, be further instances of definitional attributions which have *not* been identified (false negatives). What is important for the evaluation of our computational approach as a whole, however, is whether it is possible to gain philosophically valuable insights based on the instances which *have been identified*.

TABLE 8 Extract of final dataframe, nlp-comparison-notebook, step 1.

	Suárez	Descartes	Ockham
1	The soul does indeed suppose a complete substance	The soul always receives them from the things that are represented by them	The intellective soul is an incorruptible form
2	The soul, however, is that which gives the being of a living thing	The soul is united to all the parts of the body conjointly	The soul is an indivisible form
3	The soul is an informing substantial form	The soul is really joined to the whole body	It exists as a whole soul in every part
4	The soul is the principle of life	The soul is of such a nature that is has no relation to extension	The soul exists in each part of the body and that
…	…	…	…

between two scholastic traditions—with Ockham and Descartes on one side and Suárez on the other—which were still the subject to philosophical debate in the early modern era.[33] The affirmation of these two chains of thought may then, according to Perler's argumentation, be used to reject the traditional 'turning point' conception of the transition between medieval and early modern philosophy with Descartes as its main protagonist in favour of a progressive, evolutionary interpretation.

Now to find out if the computational approach works, the decisive question is whether an in-depth analysis of the final dataframe *alone* may lead to the same insights than those of Perler's in-depth analysis of *all* the relevant writings. According to Perler's traditional analysis, Descartes implicitly rejects two specific, closely connected assumptions in Suárez's account on the soul: the multiplicity thesis and the divisibility thesis (Perler, 2013, pp. 23 and 24, respectively).[34] The first thesis takes the soul to be essentially complex: different kinds of activities of the soul are assumed to be distributed among multiple faculties, all of which are considered to be distinct from one another. Perler then takes the second thesis to follow directly from the first one in assuming that if the faculties are distinct from each other, then the soul must be divisible, with each faculty having its own, more or less autonomous, principle of operation (Perler, 2013, p. 29). To Descartes, on the other hand, Perler ascribes a reductionist and dualist approach to the soul—which he traces back to Ockham (Perler, 2013, p. 32f.).

While the former author is said (i) to assume a dualism between soul and body, and (ii) to conceive the soul as being essentially simple by reducing its two faculties to one and the same soul, the latter assumes a dualism between the rational and the sensory soul, but (ii) reduces the two souls to one and the same principle of operation. Perler therefore concludes that Descartes' account of the soul is not as innovative as it may seem and subsequently infers the consequences outlined above with respect to the conceptual continuity between medieval and early modern philosophy (Perler, 2013, p. 37f.).

Table 9 below shows a selection of Descartes' and Ockham's definitional attributions gathered in the final dataframe that turned out to be particularly relevant for evaluating the conceptual continuity between these two authors.

Descartes' definitional attributions in lines 4, 7, 17, and 18 do indicate a dualist conception of the relation between soul and body: the author locates the soul in the 'small gland' of the

[33]These two lines of thought correspond to the operationalisation of conceptual continuity and innovation introduced above, which, in turn, draws on Perler's approach.

[34]Perler also ascribes a third thesis to Suárez, the inaccessibility thesis (Perler, 2013, p. 24), which does, however, not play a major role in his subsequent analysis of the conceptual (dis)continuities between Suárez, Ockham, and Descartes. Therefore, it shall not be subject of discussion in this section.

TABLE 9 Selection of definitional attributions in Descartes' and Ockham's writings; the line numbering corresponds to the original dataframe produced by the nlp-comparison-notebook.

	Descartes	Ockham	
2	The soul is united to all the parts of the body conjointly	The soul is an indivisible form	2
3	The soul is really joined to the whole body	It exists as a whole in the whole body and as a whole in each part	6
4	The soul is of such a nature that it has no relation to extension	The soul in the head is not distant from the soul in the foot	12
7	The soul has its principal seat in the small gland located in the middle of the brain	The two souls are distinct	15
11	This soul has within it no diversity of parts	The sentient soul is extended and material the intellective soul is not	16 17
17	The soul is immediately advised about things that harm the body	The sentient soul does not remain after the separation of the intellective soul	20
18	The soul is immediately advised about things useful to the body only through some sort of titillation	The sentient soul is introduced temporally before the intellective soul	21
22	The soul not only understands and imagines but also has sensory awareness	The soul can move the body through its organic powers	24
		The soul is indeed able to will—and perhaps it in fact wills	35

brain (line 7) and explains the interaction between soul and body 'through some sort of titillation' (18), while at the same time asserting that the soul has 'no relation to extension' (4)—as opposed to the body. Analogously to Descartes' soul-body dualism, indications for Ockham's soul-soul dualism can be found in lines 15, 16, 17, 20, and 21: he conceives the sentient and the intellective soul to be distinct from one another (15) and explicitly attributes to them opposing qualities with respect to their materiality and extension (16 and 17). In addition, while he conceives the sentient soul to be introduced 'temporally before the intellective soul' (21), he claims that the former 'does not remain after the separation' of the latter (20). But in spite of the strict dualist distinction between the *two souls*, he attributes the 'organic powers' to move the body (25) and the ability 'to will' (35) to *the* soul—which he further claims to exist 'as a whole in the whole body and as a whole in each part' (6, also see 12). Assertions like these thus point to the reductionist account that Perler attributes to Ockham's conception of the soul. Descartes' definitional attributions in lines 2, 3, 11, and 22 do likewise indicate a reductionist account with respect to the different faculties of the soul: while the soul 'not only understands and imagines but also has sensory awareness' (22), he claims that the soul is 'united to all parts of the body conjointly' (2, also see 3) and has 'within it no diversity of parts' (11). Hence, the analysis of the definitional attributions alone and Perler's traditional analysis of the entire literature do indeed allow to draw the same conclusion: both Descartes and Ockham conceive of the soul in reductionist and dualist terms. Their accounts of the soul can therefore be seen as conceptually continuous.

Concerning Perler's second claim, that is, the conceptual discontinuity between Suárez and Descartes, the picture drawn by the set of definitional attributions gathered in the course of the investigation is a bit different (Table 10).

On the one hand, there are some definitional attributions which point to Perler's characterisation of Suárez's conception of the soul as being essentially complex. In accordance with Perler's analysis, Suárez assigns different tasks and responsibilities to different parts of the

TABLE 10 Selection of definitional attributions in Suárez's writings; the line numbering corresponds to the original dataframe produced by the nlp-comparison-notebook.

	Suárez
8	The soul is the first principle of all perfections and operations of a living thing as such
15	The soul is not always separable from its operation
17	The vegetative soul, which is more imperfect
19	The vegetative soul is always subject to its own operation
22	The soul is the principle of activities which living things are able to have in themselves
37	The soul has a great diversity in its works
40	The whole human soul existing in each part informs them in diverse ways
41	The soul requires diverse dispositions in diverse parts
92	The rational soul is the form of the body

soul: in maintaining that the soul is 'not always separable from its operation' (15), he asserts that it is only the *rational* soul which is 'the form of the body' (92), while the *vegetative* soul, which is 'more imperfect' (17), is 'always subject to its own operation' (19). But asserting that the soul has 'a great diversity in its works' (37) for which it 'requires diverse dispositions in diverse parts' (41) does not yet imply that each part of the soul must also have its *own* principle of operation—as, for example, opposed to Ockham's reductionist account. In other words, the multiplicity of *operations* of the soul does not necessarily imply a multiplicity of *principles* of operations which all function more or less autonomously. So while these definitional attributions support the multiplicity thesis that Perler attributes to Suárez's account of the soul, they do not confirm the divisibility thesis. Suárez's assertion that the soul is 'the first principle of all perfections and operations of a living thing' (8, also see 22 and 40) may even indicate a reductionist strategy: even if the different parts or faculties of the soul may come with their own realm of operations (multiplicity), this passage seems to indicate that there is *one* overarching principle of operations, namely the soul as a whole (reduction). Admittedly, neither Descartes nor Ockham speak of *first* principles—a terminology that may indeed suggest the existence of 'secondary' principles of operation in Suárez's account. And even though Perler acknowledges this 'hierarchical conception of the soul' (Perler, 2013, p. 14.), he draws on the author's assertion that the different parts of the soul form 'a unity' by aggregation only (Perler, 2013, p. 17) to conclude, nonetheless, that Suárez conceives of the soul as being essentially divisible (Perler, 2013, p. 24). So while Perler takes multiplicity and divisibility to go hand in hand and thereby rules out the possibility of interpreting Suárez as reducing the multiplicity of (principles of) operations to one 'first principle', the present analysis raises the question if Suárez's commitment to multiplicity really is incompatible with a reductionist interpretation of his approach to the soul.

To sum up, the results of this analysis are in line with Perler's claim regarding the conceptual continuity between Ockham and Descartes and, by implication, between medieval and early modern philosophy. The affirmation of the multiplicity thesis makes Suárez's account already less conceptually continuous to either Descartes or Ockham than the observed continuity between these two. Whether the multiplicity thesis is compatible with a reductionist interpretation and how it relates to the divisibility of the soul could not be conclusively answered on the basis of the definitional attributions gathered in the synoptic presentation. As noted above, the present application was first and foremost an experiment. Some texts, however, did not meet the two criteria for a computational analysis at the time the experiment

was conducted. Unless *all* (relevant) writings have been taken into account, the picture remains incomplete. And it goes without saying that unless the *original* writings are used, no conclusive judgement can be made.

4 | DISCUSSION

Having explored the computational approach in the previous section, it is now time for a critical evaluation. We will start by highlighting the strengths and prospects of the approach as a whole. We will then turn to the challenges encountered during its application and discuss how they illustrate what we can and cannot expect from the use of specialised NLP-models in this area of research.

4.1 | Strengths and prospects

The approach developed in here demonstrates how to combine the advantages of a computational in-breadth analysis with the merits of a traditional in-depth analysis in a philosophically fruitful way. Unlike traditional scholarship's selective approach to historical sources, no choice whatsoever must be made for reducing the amount of literature to a humanely manageable size. Instead, the systematic approach of the current method allows to take into account all the writings of all the authors that may be relevant for a given research question, at least in principle. From a purely practical point of view, the experiment presented in the previous section was more selective than traditional scholarship since not every writing included by Perler met the two criteria for computational analysis at the time the experiment was conducted. But in this context, we must distinguish between two different reasons for this kind of selectivity, that is, the mere number of sources included:

(a) *Selectivity due to limited cognitive resources*: human scholars cannot include all the historical sources that may be relevant for a given research question because the number of potentially relevant sources often exceeds what a scholar could read within the timeframe of a given research project.
(b) *Selectivity due to non-accessibility*: computational analysis cannot include historical sources unless they are accessible in digital form and in a natural language for which a suitable NLP-library has been developed.

The former reason (a) entails selectivity that is inevitable: we cannot overcome the limitations of our cognitive resources. The latter reasons (b), however, only entails selectivity due to current practical constraints: as soon as a historical source becomes accessible in digital form and in a natural language for which a suitable NLP-library has been developed, it can be processed by computational analysis. Since the time the experiment was conducted, significant advances in text recognition and NLP of ancient languages have been made that now allow access to almost all available historical sources.[35] Therefore, selectivity due to non-accessibility is hardly relevant for computational methods today.

By drawing on the in this area of research yet unexplored potential of modern machine learning techniques for the computational analysis of philosophical literature, we have shown that the use of specialised NLP-models could enable scholars to identify all those text passages

[35]Some problems of selectivity may nonetheless remain, for example, due to legal or conservation restrictions. These apply to both humans and machines. In some cases, however, machine processing of text fragments even allows better accessibility than for the traditional scholar, as with the burnt papyri of Heraklion.

in which terminological and conceptual shifts between the ways different authors characterise a philosophical concept may potentially be detected. Using these models for research in the history of ideas is thus not intended to replace the traditional in-depth analysis, but to make better use of it. They are not only key to overcoming the selectivity of the included sources by making the research process much faster, but they also address a number of other problems that human scholars face, such as remembering details and sometimes only later becoming aware of subtle aspects or certain passages in the text that are relevant to their argument. This cognitive selectivity *in* the included sources affects scholars' ability to make complex interpretations and is distinct from the selectivity *of* the included sources imposed by the time required for the physical act of reading itself. In contrast, computational analysis is not affected by this cognitive selectivity—at least not directly, as we will see in the next subsection.

And even if human scholars could take into account all relevant text passages, there is a significant limitation in the justification of interpretive hypotheses compared to computational methods. This limitation concerns the recipient and the factual justification of the hypotheses. In philosophical papers, not all the evidence is usually cited, but only selected examples are quoted or indirectly referenced. As a result, the historical evidence as a whole remains intransparent and the comprehensibility of the argumentation is limited. Computational methods can overcome this limitation by making the overall evidence of the text passages comprehensibly verifiable for the reader—as in the case of the synoptic presentation of all instances of definitional attributions that is included in the supplementary material of this paper. In this way, computational methods can make the justification of an interpretive hypothesis transparent in relation to the overall evidence. The importance of this additional dimension of historical and philosophical reasoning is likely to increase rapidly in the humanities, as computational methods open up new possibilities for improving the quality of reasoning and ensuring the verifiability of hypotheses.

But the approach presented in this paper also enables us to shed light on potential new problems that arise from using computational methods in this area of research. These problems will be discussed in the next subsection.

4.2 | Limits of current machine learning techniques

Even though the specialised NLP-model for text classification reached a near-perfect accuracy in predicting instances of third-party ascriptions in the training sessions conducted with Prodigy, there were considerable differences in determining the exact threshold for reliably excluding such instances depending on each author's particular style of writing to which it was applied. A more consistent threshold that decreases the need for human intervention could probably be achieved by training the model on more data.

And while our approach demonstrates how some problems of cognitive selectivity *in* the sources could be solved with the help of computational means, the use of specialised NLP-models also leads to a new, somewhat related problem: the elimination of potentially relevant textual context. Our experiment illustrates this very well. After a careful inspection of the dataframe, it turned out that it was not always apparent on the level of single sentences whether they really express the final viewpoint of the respective author. While the model for text classification proved—despite the threshold problem—to be rather efficient in detecting cases of third-party ascription whenever there *was* an indication in the particular sentence, there were also cases in which no such indication was given *in* the sentence but which nevertheless turned out to belong to this class as soon as one takes a look at the broader context. In the worst case, this resulted in identifying two contradictory statements as instances of definitional attributions—as, for example, in the case of Suárez (Table 11).

TABLE 11 Suárez, extract of dataframe without false positives, nlp-analysis-notebook, step 4; line numbering corresponds to the original dataframe.

	def_attrib	Meta	Text	tp_ascrip
…	…	…	…	…
85	The human soul is material	{'source': 'Suarez, De Anima'}	Therefore the human soul is material.	0.002473
86	The human soul is immaterial and subsistent	{'source': 'Suarez, De Anima'}	Let the first conclusion be as follows: It is evident by natural reason that the human soul is immaterial and subsistent.	0.002355
…	…	…	…	…

A brief glance at the broader context reveals that the assertion in line 85 is part of a detailed explanation of the extent to which the soul *might be called* material, whereas the latter statement affirms that the soul is *in fact* immaterial (Suárez, 2011, pp. 112 and 113 respectively). Although these cases were rare, it is a serious problem which one has to be aware of when interpreting the final results of the computational analysis.[36] The lack of textual context, which causes this problem, could possibly be circumvented by applying the model for text classification to entire paragraphs rather than to individual sentences. But as this would, in return, bear the risk of missing out actual instances of definitional attributions that may appear in a paragraph mainly composed of third-party ascriptions, it was decided not to adopt this strategy for the present approach. Future research, however, might consider running their text classification model twice on every text: first on the level of entire paragraphs and then a second time on the level of individual sentences. Such an approach could facilitate the identification of third-party ascriptions even when there are no indications on the sentence-level, while avoiding the shortcomings of classifying paragraphs only. These considerations make it clear that the current use of NLP-based text classification is not yet completely free from error. But having raised awareness of these challenges through the present investigation, we believe that it will become much easier for future research to address them.

The performance of the NLP-model for named entity recognition suggests that its potential for the intended purpose is much less promising. As already observed during its development, the model's accuracy in correctly predicting instances of definitional attributions is not yet entirely satisfactory. In conjunction with the manual identification of false positives, it nonetheless turned out to be capable of producing useful results. But due to considerable differences in the length of the entities that had to be recognised—partly caused by the variety of characterisations that were intended to be captured under the label of definitional attribution—, it seems questionable whether NER is really the appropriate tool for this task. A possible way of enhancing the model's accuracy could consist in dividing the entity class into different subcategories. The more specific, that is, the more fine-grained, the categories for NER, the easier it will become to make consistent annotation decisions. And the more consistent the decisions, the higher the expected accuracy of the resulting model. But even if one breaks down a given entity class into simpler subcategories, a more general problem with the development of specialised NER for this area of research will probably persist. Both the creation of an annotation policy and its application to the data require clear-cut decisions as to

[36]Since this is not a case of a false positive in a strict sense—as there is nothing wrong with the prediction—, it has not been eliminated at the end of the nlp-analysis-notebook. However, it was not further taken into account for the analysis of conceptual continuity. The same goes for Ockham's assertion that *in a certain sense*, the soul in the head *is* distant from the soul in the foot, although they are in fact *not* distant (lines 15 and 16 of the dataframe produced by the corresponding nlp-analysis-notebook, step 4).

what counts as an instance of the named entity in question—and what does not. When analysing philosophical literature, one is, however, often faced with cases for which no such completely unequivocal decision can be made.[37] Even with a well-conceived annotation policy, it turned out that some interpretation of the previously established rules is often required. The tension between the clearcut decisions that must be made for creating the training data and the occurrence of ambiguous cases which do not readily fit presents a serious obstacle to the machine's learning process and, by implication, to the success of the entire endeavour. A frequent occurrence of such ambiguous cases should therefore be taken as an indication that one is going beyond what can be expected from current machine-learning techniques.

In the present case, the difficult annotation decisions, the relatively low F-score, and the high proportion of false positives are all different dimensions of the same problem. Taken together, they suggest that the identification of definitional attributions may well be a task too complex for NLP-based named entity recognition. Unlike the challenges identified for text classification, they clearly point out the limits of the current means of NLP and machine learning in this area of research.

Furthermore, developing and applying the NER-model revealed another important problem that arises from the NLP-based analysis of historical sources. The crucial piece of evidence that Perler draws on (referred to above) to ascribe the divisibility thesis to Suárez despite his hierarchical conception of the soul is contained in a text passage in which it is not the soul itself but its faculties which are being characterised (Perler, 2013, p. 17). While it seemed reasonable to exclude such genitive constructions when developing our annotation policy for definitional attributions, this particular genitive construction eventually turned out to be highly relevant for the research question. This is not a further argument against the performance of NER-models for the analysis of philosophical literature. For there is nothing wrong with the model's failure to detect this passage. Rather, it casts doubt on whether the exploration of philosophical research questions can be based solely on a computationally generated list of instances of particular NER-categories that are as complex as our category of 'definitional attributions'. For if we do, we can hardly rule out the possibility that through our annotation policy, the NER-category overlooks an important piece of evidence, even if it was specifically designed to detect such evidence.

5 | CONCLUSION

It should be clear from what has been said that the specific computational approach chosen for the purposes of this paper has its own limitations. But having shown in Section 3 that the approach is capable of producing philosophically valuable results, we have illustrated how NLP and machine learning can be used to overcome important limitations of traditional scholarship: instead of restricting investigations of conceptual continuities and innovations to the analysis of only a few selected authors and writings, the method developed in here shows that computational means can help us to consider every source that might be of interest for a given research question. By supplementing the traditional methodological toolbox with the systematic tools of computational in-breadth analysis, it may thus become possible to make more complex interpretations and to make the justification of an interpretive hypothesis more transparent for the reader.

The current project also brought to light how far one can go with the most recent advances in NLP and machine learning. We thereby discovered important limits to what one can

[37]For the NER-model developed in here, one may, for example, object to our decision to *include* definitional attributions by negation but to *exclude* negated definitional attributions. After all, a negated definitional attribution may also tell us something about the author's view on the concept at stake.

reasonably expect from these tools at present. While the NLP-based text classification already proved to be a highly useful method for the computational analysis of philosophical literature, our experiment showed that it is not only very difficult for machine learning to clearly delimit what does and what does not count as an instance of a definitional attribution. It also illustrates that even though the computational process itself is not affected by cognitive selectivity, it can still be indirectly affected because its development requires human intervention: in the present case, we only later discovered the relevance of certain passages in the text that we thought reasonable to excluded due to their grammatical structure when conceiving our annotation policy. The question, then, is not whether machines or human scholars produce better results, but how to combine the virtues of both in a way that minimises the potential weaknesses of the other. Future research on named entity recognition could, for example, try to focus on less complex tasks than the one envisaged in the current project that are less likely to be affected by the problems arising from cognitive selectivity during their development. Through our transparent documentation of all the challenges encountered during development and application, we hope to have provided valuable insights that will guide further advances in this field. Our investigation shows that the use of specialised NLP-models could be a promising extension to the methodological toolbox of traditional scholarship, provided that the new problems we have identified in this paper are kept in mind.

AUTHOR CONTRIBUTIONS

Simon Brausch and Gerd Graßhoff conceived the experiment. Simon Brausch designed and performed the experiment, analysed the data, and wrote the manuscript. Simon Brausch and Gerd Graßhoff revised the manuscript.

ACKNOWLEDGEMENTS

We are extremely grateful to Pierre-Louis Dubouilh for discussing numerous technical questions and without whom SB would probably not have been able to design the experiment on which this paper is based. We would also like to thank the participants of a research colloquium at the Humboldt-Universität zu Berlin (HU) for a stimulating discussion about this project during its development as well as the two anonymous reviewers for their valuable comments on the manuscript. The experiment was conducted while SB was a Master student at the HU, and the manuscript was written and revised while he was a member of the Max Planck Research Group (MPRG) "Practices of Validation in the Biomedical Sciences" at the Max Planck Institute for the History of Science (MPIWG), Berlin. We extend our thanks to both the MPRG and the MPIWG for their generous support. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

The five Jupyter Notebook documents that form the basis of the computational approach developed in this study as well as the results of the computational analysis discussed in Section 3 are included in the supporting material for this paper. The two specialised NLP-models for text classification and named entity recognition referred to in Section 2 are available from the corresponding author upon request.

ORCID


Simon Brausch http://orcid.org/0000-0002-7423-1907

REFERENCES

Aiello, K.D. & Simeone, M. (2019) Triangulation of history using textual data. *Isis*, 110(3), 522–537.

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*. Palgrave Macmillan, pp. 85–101.

Andow, J. (2015) How distinctive is philosophers' intuition talk? *Metaphilosophy*, 46(45), 515–538.

Betti, A., Van Den Berg, H., Oortwijn, Y. & Treijtel, C. (2019) History of philosophy in ones and zeros. In: Curtis, M. & Fischer, E. (Eds.) *Methodological advances in experimental philosophy*. Bloomsbury Publishing, pp. 295–332.

Descartes, R. (1985) *The philosophical writings of Descartes*, vol. 1 & 2 (Trans. by Cottingham, J., Stoothoff, R. & Murdoch, D.). Cambridge University Press.

Gibson, A. & Ermus, C. (2019) The history of science and the science of history: computational methods, algorithms, and the future of the field. *Isis*, 110(3), 555–566.

Laubichler, M.D., Maienschein, J. & Renn, J. (2013) Computational perspectives in the history of science. *Isis*, 104, 119–130.

Ockham, W. (1991) *Quodlibetal questions*, vol. 1 & 2 (Trans. by Freddoso, A. J. & Kelley, F. E.). Yale University Press.

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

Perler, D. (2013) What are faculties of the soul? Descartes and his scholastic background. In: Marenbon, J. (Ed.) *Continuity and innovation in medieval and modern philosophy: knowledge, mind and language*. Oxford University Press, pp. 9–38.

Suárez, F. (2011) *Selections from De Anima* (Trans. by Kronen, J. & Reedy, J.). Philosophia Verlag.

Van Wierst, P., Vrijenhoek, S., Schlobach, S. & Betti, A. (2016) Phil@Scale: computational methods within philosophy, In: *CEUR Workshop Proceedings*, 1681.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Brausch, S. & Graßhoff, G. (2023) Machine learning for the history of ideas. *Future Humanities*, 1, e6. https://doi.org/10.1002/fhu2.6