DIGITAL CLASSICS ONLINE

Measuring Philosophy in the First Thousand Years of Greek Literature

Thomas Koentges

Abstract: In this pilot study, the author applied Latent Dirichlet Allocation (LDA) topic modelling to train a machine to automatically identify philosophical passages in a corpus (produced by the Open Greek and Latin group and the Perseus Digital Library) representing the preponderance of extant works of the first thousand years of Greek literature. The author utilised qualitative data analysis, document vectors produced through topic modelling, and Classics domain knowledge to distil three numeric scores for philosophical text in Ancient Greek. One measures "good and virtue", while the second score measures "scientific inquiry", and the third combines the two to measure "philosophicalness". The scores are applicable to passages, works, and workgroups within the corpus and not only performed well in identifying works by philosophers but also identified passages by non-philosophers containing philosophical components. Thus, this method represents a widely applicable, scalable, and unbiased way to find research-relevant passages in a corpus that is too large to be read in its entirety.

1. Introduction[1]

The research on which this paper is based exists in an interdisciplinary nexus of academic disciplines and research fields. It touches on text-reuse and genre detection, classical philology, and natural language processing (NLP). While topic modelling and automatic genre detection are wide fields that utilise a multitude of different computational methods, in what follows, the author utilises topic modelling as a means to an end, by producing a score that can detect structural and content-based genre signals, with a special focus on Ancient Greek philosophy.

Genre definitions have long been discussed in the Classics community. Published discussions mainly focus on sub-genres,[2] while the difficulty of tracing and defining genre diachronically and synchronically has been highlighted.[3] Leaving this discussion to the Classics community, for this paper, it is enough that we agree (or at the very least keep an open mind) that there is something that can be called "genre" and that when deciding to which genre a text belongs objective and observable criteria can be used. Following this empirical stance, the article is focused on the computational implementation of a score that can measure the entity "genre" with a focus on the genre "philosophy".

While the public perception of Greek philosophy is tightly connected with both Socrates and the writings of Aristotle and Plato in the fifth and fourth century BCE, and in Classics, there has been extensive

1 The author would like to gratefully acknowledge Harvard's Center for Hellenic Studies for the Residential Fall Fellowship (2017) and on-going appointment as Fellow in Historical Language Processing and Data Analysis (2018–) that made this research possible. Thanks too to my colleagues on the Open Greek and Latin project (OGL) team for their commitment to the openly accessible data on which this research is based.

2 See, for instance, Nagy (1994) or Van Dijk (1997).

3 See, for instance, Calame (1974) and Rotstein (2012).

research on the pre-Socratics as well as the neo and middle Platonists, beyond this scope, structured research of Greek philosophy in a wider sense is often more difficult because of the vastness of Greek literature. Passages containing philosophical content are difficult to trace if they do not contain verbatim or almost verbatim references to other philosophical text. This paper attempts to address the challenge of finding research-relevant passages in vast corpora.

One approach to solving this challenge is to find all potential paraphrases using word vectors and then assess the so-called philosophicalness of a paraphrase through qualitative methods, as in the Platon Paraphrasen Digital (PPD) project.[4] In this project, word vectors are used to calculate the Word Mover's Distance of a passage to all passages in a corpus using a brute-force implementation.[5] The top 100 passages can then be qualitatively evaluated by philologists using traditional methods. Since one of the main goals of the PPD project is to detect references to Plato in large corpora, focussing on word vectors is sensible. However, this article is intended to offer a complementary approach to this project by using the quantitative method Latent Dirichlet Allocation (LDA) topic modelling[6] to produce document vectors—and only indirectly word vectors—that are based on the structural and thematic content of a passage. Those document vectors will be assessed through a mix of quantitative and qualitative methods to try and find a quantitative score that can be associated with a direct measure for philosophicalness.

The choice of topic modelling for this task builds on recent research in Digital Humanities. While initial genre-detection work more often applied stylometric measures to literary corpora,[7] topic modelling for genre detection has been widely applied in other fields; for instance, to French literature by Schöch (2017). In his paper Schöch demonstrates convincingly that topic modelling can detect a genre signal, although he focuses on a more specific corpus and utilises only three genre categories: Comédie, Tragi-comédie, and Tragédie. This paper expands on this precedent by including close to 30 million words and by employing bibliographic metadata for 29 genres, albeit that the test for philosophy is essentially binary. It also has a different language focus, Ancient Greek, a language with a much more complex morphology.[8] While it has already been shown that this morphological complexity signifies a challenge for topic modelling Ancient Greek corpora,[9] the size of the research corpus employed here mitigates that challenge and thereby this is the first large-scale computational analysis of genre in the domain of Ancient Greek literature.

In this pilot study, the author applied topic modelling to train a machine to automatically identify philosophical passages in the First1KGreek corpus, a corpus produced by the Open Greek and Latin group (OGL) containing, as its name suggests, the first 1,000 years of Greek literature.[10] Using correlation analysis to compare topics generated by an LDA model helped to distil a numeric measure for philosophical text in Ancient Greek, thereby enabling the detection of passages with either philosophical content or structure. This was possible through the generation of three scores: two measure structural and thematic aspects of Ancient Greek philosophy, while a third is a combination of those two scores (and thus resembles a single score) that helps to rank passages based on their philosophicalness. In combination,

4 See https://digital-plato.org [Last access 10-06-2020]. The paraphrasis finder can be found at https://paraphrasis.org [Last access 10-06-2020]. See also some earlier attempts in Geßner (2010). For an overview of the project and methods employed see also Schubert et al. (2019).

5 Pöckelmann et al. (2017). For the Word Mover's Distance see Kusner et al. (2015).

6 Blei et al. (2003).

7 See, for instance, Stamatatos et al. (2000) or Jockers (2013), 75–77.

8 Dik / Whaling (2008).

9 By Koentges (2016a) and Wishart / Prokopidis (2017).

10 Crane et al. (2020).

these three scores not only worked well to identify works written by philosophers, but also to extract individual passages from works by non-philosophers that nonetheless contain philosophical arguments.

This article therefore argues that using topic modelling to generate vectorisation of the documents represents a widely applicable, scalable, and unbiased way to find research-relevant passages in a corpus that is too large to be read in its entirety. The article begins by describing the corpus and introducing historical language processing (HLP) in general, before outlining the method used to produce the topic model and explaining the development of the score and how it was tested. It then offers some examples of how the score can be applied to the different levels of granularity within the research corpus, potentially expanding the number of works and passages taken into consideration when researching Greek philosophy beyond those already well known. By doing so, and by sharing the code on which the research is based, it is hoped that similar methods can be applied to multiple research areas and genres, potentially going some way to address domain bias in research.

2. The Corpus

The research described in this paper is based on what, for simplicity's sake, can simply be known as the First1KGreek corpus. The origins of the corpus reach back to 1987, when the Perseus Digital Library made the first digitisations of Ancient Greek text. The corpus has, however, grown and moved to employ open data principles since then, and is now a collaboration between several universities and research institutions, including the University of Leipzig, Mt Allison University in Canada, Harvard's Center for Hellenic Studies, Tufts University, and Virginia University. These organisations, including Perseus, have been joining forces through the Open Greek and Latin group and together are trying to make available under an open licence every extant work of the first 1,000 years of Greek literature. For this research, the OGL corpus was supplemented with the CTS-compliant TEI XML files of the Perseus Digital Library.[11] All texts have been transferred from TEI XML to a simple text format using the TEItoCEX application written by the author to aid transferability and analysis of the OGL and Perseus CTS corpora.[12]

In its digitisation process, OGL employs a data-centric approach: the underlying data structure of the corpus is always accessible and there is no single black-box database. While OGL uses databases in the background, the underlying data structure is also pushed to open platforms, such as Zenodo.[13] The OGL group also encourages direct interaction with our data: the corpus is directly downloadable and new releases are frequently published. In addition, researchers can also precisely access the exact texts they want by utilising the Canonical Text Services and Collections, Indices, Texts, Extensions (CTS/CITE) framework that makes the corpus linked-data-ready.

CTS/CITE is a highly precise framework to reference research data in textual humanities in a machine-actionable way.[14] Traditional reference systems rely on a good deal of human interpretation: for instance, traditionally Plato is often cited according to the pages in the edition of the Corpus Platonicum published by Henri Estienne (Stephanus) in 1578. Yet, if one looks at different modern editions of Plato's text, where a certain paragraph or page starts can differ from edition to edition. Because the original Stephanus paragraphs often break the text mid-sentence and sometimes even mid-word, comparing the text of two different editions needs some adjustments and interpretation by a human reader. With CTS/CITE we normalise such phenomena, while also adding granularity to the text. Through the CTS/CITE

11 See https://github.com/PerseusDL/canonical-greekLit [Last access 10-06-2020].

12 Koentges (2020c).

13 Crane et al. (2020).

14 For an accessible introduction see Koentges et al. (2020).

framework we can therefore access different levels of information, with most of the CTS passage/verse nodes coming in at under 100 words. This means that a machine can process not only all workgroups in the First1KGreek corpus, but also all works, and even all passages within those works.

The First1KGreek corpus comprises 174 workgroups. A workgroup is a metadata category that includes all works that have a common origin or author. For instance, the New Testament and the *Corpus Platonicum* are workgroups and so often workgroups can be associated with a single author. Within those workgroups, the First1KGreek corpus comprises 1,044 individual works (at the time of writing). Figure 1 shows those individual works depicted as bubbles, with bubbles of the same colour belonging to the same workgroup and a bubble's size reflecting the work's word count. In figure 2 we see the same visualisation, but with the individual CTS passages as tiny points. Hence, it's a question of granularity, with division into individual works (Fig. 1) then into individual CTS nodes, which are often paragraphs, sentences, or lines (Fig. 2).

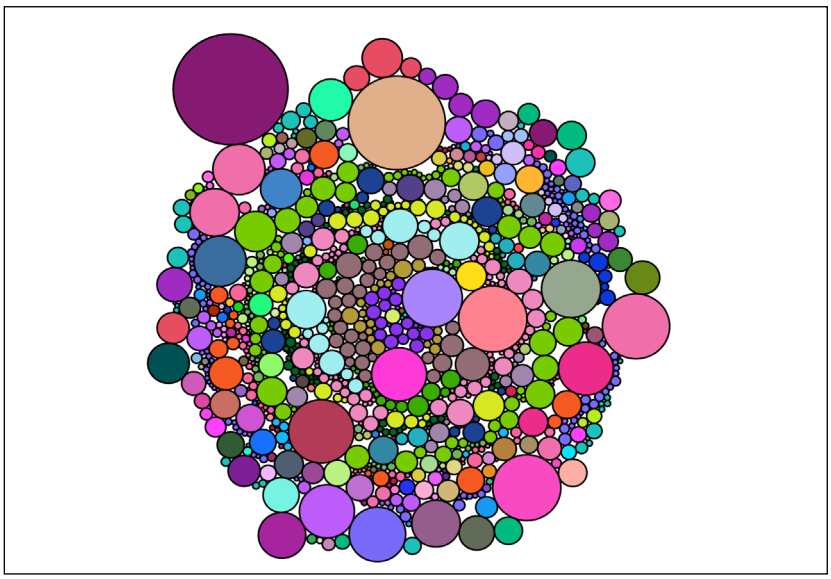

Fig. 1: First1KGreek works as bubbles.[15]

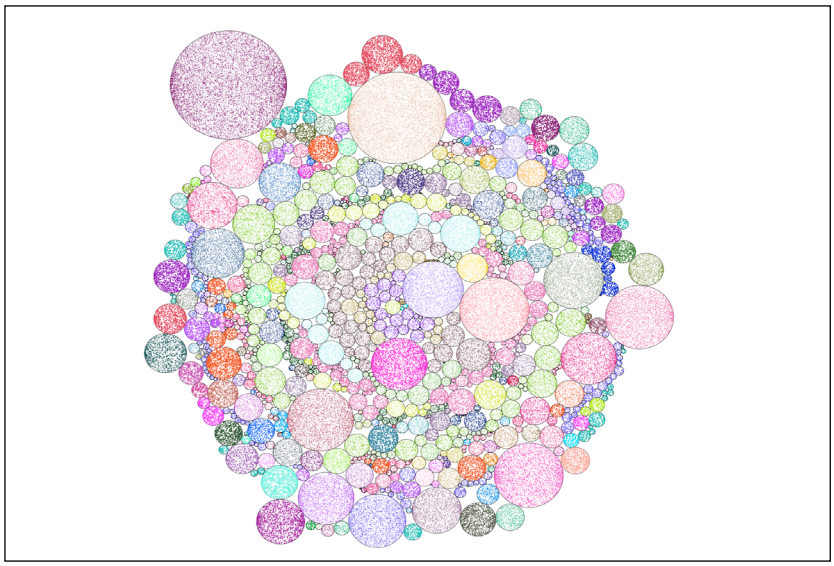

Fig. 2: First1KGreek corpus depicted as bubbles (works) and points (CTS passage nodes).

15 All visualisations in this paper were produced in the programming language R. For complete transparency the source code for visualisations and data exploration, as well as additional visualisations, are also published on Zenodo, see Koentges (2020e).

At the time of writing, the First1KGreek corpus has over 200,000 CTS nodes. The research described in this paper utilises this granularity by analysing the individual nodes and then using the results and higher-level metadata to develop a score that can then be reapplied at the passage level. It is, therefore, the first research to employ the vast First1K Greek corpus at this very granular level.

Put another way, those 200,000 nodes represent the just-under 30 million words of Ancient Greek that make up the First1KGreek corpus. It is thus fair to say that this is beyond an individual scholar's comprehension if they merely rely on their memory, and it is also likely more than any modern scholar has read. Yet, if we want to study these texts but also want to ensure that we are not biased by our personal selection (which is grounded in our own interests, education system, and research trends) and instead want to access text and information that we did not know through those channels, then it is important to be able to delve into the corpus with NLP methods.

3. Historical Language Processing

NLP refers to the understanding and creation of human languages by a machine. However, as Neel Smith has pointed out, NLP for historical languages is tricky because of their often-higher morphological complexity.[16] Yet, this does not mean that modern languages are easier in general; rather, complexity simply changes over time. For instance, English is less morphologically complex than Ancient Greek, but one could argue that it has a much stricter word order and a higher idiomatic complexity. These different complexities can influence how successful NLP methods can be, because most of those methods include the same central principle.

One of the most widely employed concepts within NLP research is the bag-of-words (BOW) principle, which in fact emerged nearly 70 years ago.[17] This suggests that when considering the semantic meaning of a word, it is much more important with which other words a word occurs, than in which order or in which part-of-speech it appears. When looking into common NLP methods like topic modelling or word embedding vectors, one will often find that the BOW principle is an integral part of the method. This is great for a language like English, because BOW effectively addresses its specific complexities and utilises the fact that the same word is relatively easy to count. For instance, the word "go" has five forms: "go", "goes", "going", "went", "gone". However, in Ancient Greek, for instance, the word βαίνω has 26 different forms in the present active alone. The occurrence of Ancient Greek words is therefore more difficult to count from a processing point of view, and so we face a token-frequency problem.

One could address this challenge by morphological normalisation—that is, reducing the words to their dictionary form—or by simply adding more data, hoping that this makes rarer wordforms more observable. In general, it is often preferable to add more data, because morphological normalisation is not lossless. In particular, it loses information about gender and number and can make a resulting model less expressive.[18]

16 Koentges (2016a) and Smith (2016).

17 Harris (1954).

18 Another approach for topic modelling or other document-based methods could be what I call "translation bootstrapping" and "morphological bootstrapping". The former method adds a translation to all documents in a morphologically less-complex language, and the latter adds all dictionary forms of the occurring words. Both enrich the documents with additional information instead of taking information away. I am currently collaborating with Klaas Bentein to apply this principle to Ancient Greek papyri; a paper is in preparation.

In any case, researchers working with historical languages often cannot simply apply methods developed for English. Instead, it's a matter of developing strategies to address differences in tokenisation, token-frequency, or other complexities. Using NLP methods together with an awareness of the special nature of historical, often morphologically complex, languages represents a sub-field of NLP that I call historical language processing. In what follows, I will describe how I applied HLP to the First1K Greek corpus.

4. Topic Modelling

The research described in this paper relates to my wider application of historical language processing to the OGL corpus of Ancient Greek. Specifically, in 2017, as part of a residential fellowship with Harvard's Center for Hellenic Studies, I began to apply a number of stylometric measures to that corpus.[19] These initial tests resulted in some interesting findings, which later led me to question whether in fact the *Menexenus* was really written by Plato.[20] As a natural extension of this research, I wanted to focus on extracting features that could help researchers find philosophy. In theory, topic modelling seemed an appropriate method for this.

Topic modelling is "a method for finding and tracing clusters of words (called 'topics' in shorthand) in large bodies of texts".[21] A topic can be described as a recurring pattern of co-occurring words.[22] Topic models are probabilistic models that are often based on the number of topics in the corpus being assumed and fixed. The simplest and probably one of the most frequently applied topic models is LDA, the Latent Dirichlet Allocation.[23] LDA is a method of statistical inference deducing the properties of an assumed underlying Dirichlet distribution by analysing the words of a corpus. LDA is based on a simplification of how texts were created, in which the word order, for instance, does not matter. In LDA's simplified model of text creation, every document, or in our case passage, draws a topic based on a probability distribution for topics for the passage and then draws a word based on a probability distribution for words in the drawn topic. Although we know this is not the true process of text generation, LDA is highly effective, because it traces recurring clusters of co-occurring words that may reveal a lot about text-reuse, recurring topics, or the transmission of ideas through time and genres.

To produce the model, I used my programme (Meletē) Tōpan, which also means that classicists can repeat the experiments described here. Tōpan is a graphical user interface (GUI) that I have been developing since 2016. It was motivated by my desire to make topic modelling more accessible for non-coding Classics scholars, and it allows any humanist to topic model their own research corpus. It is open source and available on Zenodo.[24] In Tōpan users can create topic models by starting with an import of their data from various formats and sources, calculating and selecting stopwords, addressing morphological complexity through normalisation, and then selecting the parameters for the LDA model intuitively. Users can also review the models through visualisation and exploration of the probability distributions, which are depicted as tables.[25] Tōpan was written specifically for working with Ancient Greek and Latin, but because it has also been used by researchers in different fields, I have implemented a language-ag-

19 Koentges (2018).

20 Koentges (2020d).

21 Posner (2012).

22 Brett (2012).

23 Blei (2012).

24 Koentges (2019).

25 For a description of the original (Meletē) Tōpan see Koentges (2016b).

nostic generalisation called tidyTōpan.[26] While the original Tōpan includes the option of morphological normalisation for Latin and Ancient Greek by either sending data to the Perseus morphology service or by setting up a local installation of LatMor,[27] I instead opted to apply the "more data-strategy" outlined in the section above to address the token-frequency problem. As a rule of thumb, the more complex the language, the more data is needed. Since this model uses almost 30 million words in over 200,000 passages, we can create a meaningful model without morphological normalisation.

It should also be stressed that I was not interested in producing the "best" possible model. Rather, I wanted a model that enables me to observe and trace specific themes in a corpus by combining probabilities for a selection of topics. Topic models are evaluated by "measuring performance on a secondary task" or by using the model to complete a held-out document and testing it against the actual words in the document.[28] All evaluation methods are computationally expensive, because one needs to fully train each model before it can be tested. For instance, in the French topic-modelling research mentioned above, Schöch trained 48 different models and tested them specifically for the task of genre classification.[29] Training a similar set of models for the First1KGreek corpus would mean several weeks of computational time if run in series and several days if the task were to be parallelised, but could potentially result in only slight knowledge gain.[30] Additionally, as Schöch has already pointed out, even when attempting to find the "optimal" settings for topic modelling, "some degree of intuitive or arbitrary decisions remain".[31] Thus, I opted for a visual and human-centric review of the topic model, utilising the graphical representation of LDAvis that is included in Tōpan and my intimate knowledge of the nature of the works included in the First1KGreek corpus, selecting a model from four candidates.[32]

The topic model I chose was produced using 100 topics (k = 100) for the almost 30 million-word corpus. The model essentially creates a 100-dimensional space in which each document is represented by a sparse vector. When visualising the model, it is possible to reduce the dimensionality either through Principal Component Analysis (PCA) or t-Stochastic Neighbour Embedding (t-SNE).[33] While t-SNE generally yields good results for clustering topic modelling data, it would also be very expensive for over 200,000 vectors in a 100-dimensional space. However, the t-SNE visualisation for the individual workgroups (at just over 174 vectors) in figure 3 already indicates the model's usefulness for genre detection. This simple "eye-balling" test suggested I could continue the analysis with this model.

26 Koentges (2020a).

27 Springmann et al. (2016).

28 Wallach et al. (2009).

29 Schöch (2017), 20.

30 Wallach et al. (2009).

31 Schöch (2017), 21.

32 For LDAviz, see Sievert / Shirley (2014).

33 See Jollife (2002) for an extensive introduction to PCA and Maaten / Hinten (2008) for t-SNE.

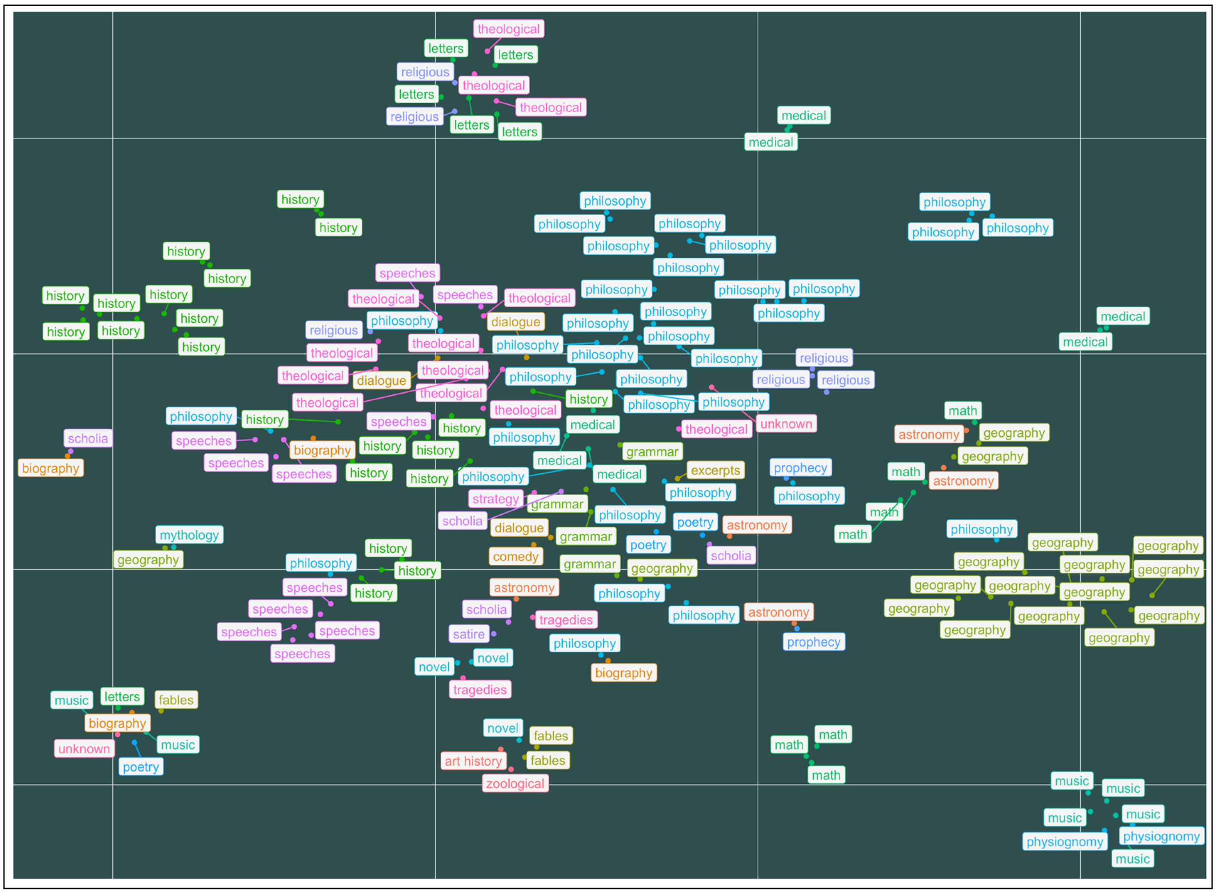

Fig. 3: t-SNE visualisation of the workgroups in the generated topic model. The points are coloured and labelled by genre based on the workgroups' metadata.

Having now described topic modelling itself, its usage in genre detection, and the selection of the LDA model used in this analysis, in the following section I will describe how I distilled a score by exploring the probability distribution of each topic within the philosophical workgroups of the corpus. I wanted to see whether I could use this topic model, which is essentially a sparse vectorization of the documents, to automatically find documents that contain philosophical text, structure, or thinking.

5. Finding the Score

To find the philosophicalness score, I followed an explorative data analysis approach by using domain knowledge to analyse how the model described in section 4 expressed relationships between workgroups and genre. I also used a mixed-methods approach when evaluating the mathematical independence of the topics within the model. To begin, I first decided to compare arguably the two most well-known philosophers of Greek literature: Plato and Aristotle. Figure 4 shows the overall distribution of topics according to the arithmetic mean of the document-topic score θ of the individual passage. Technically, the θ-value expresses the probability that a specific document belongs to a topic in a generated model. When analysing a specific topic model, the θ-value can thus be used to estimate how prevalent a topic is in a work or workgroup. I hoped that topics that describe philosophical content would be among the most prevalent for both philosophers.

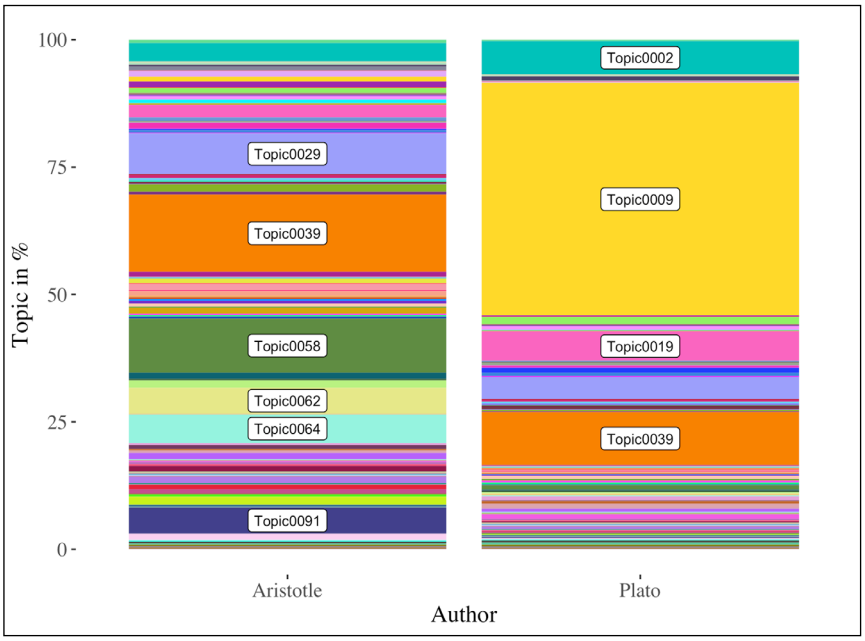

Fig. 4: Topic distribution in Aristotle and Plato. The different colours represent different topics (labels are given for the most prevalent topics).

Yet, even a cursory look at figure 4 shows that Plato and Aristotle strongly differ regarding one topic: topic 9. The content of this topic is shown in figure 5, which is a new type of topic visualisation that I have developed with the hope of moving beyond the limitations of word clouds. I used the distribution of words over the topic, but also the results of a relevance-ranking algorithm,[34] in order to put dots on the polar coordinate system with an attached label containing the matching word. The lower the radiant distance of a point to the centre of the polar coordinate system and the bigger the label, the greater its importance to that topic.

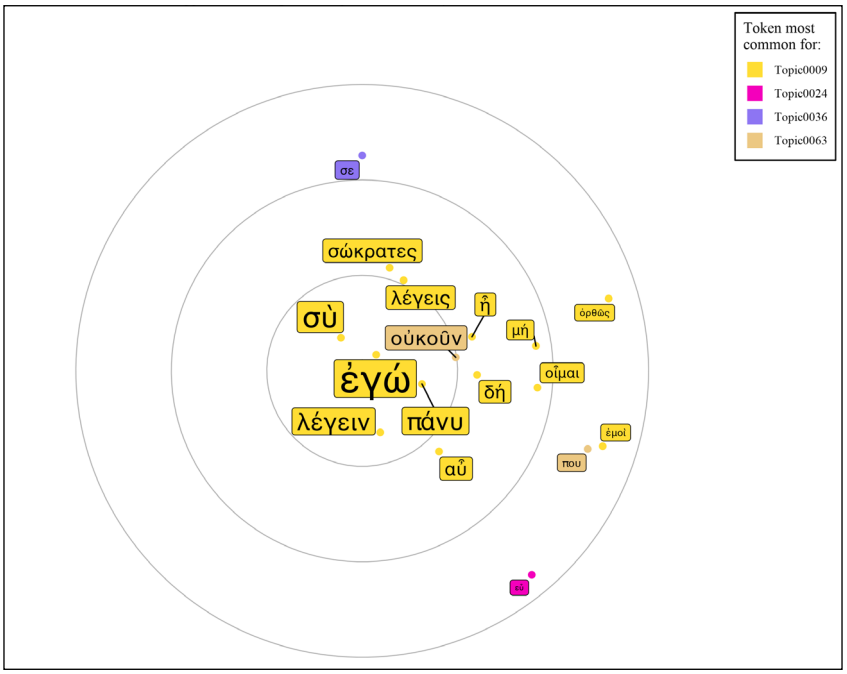

Fig. 5: Visual representation of topic 9. The closer words are to the centre, the more relevant they are for the topic.

34 For the ranking method used, see Sievert / Shirley (2014).

When one looks more closely at the words shown in figure 5, it becomes apparent that this topic is of a structural rather than thematic nature; that is, it mainly features words typical for dialogue. This corresponds to the widely known fact that the preponderance of Plato's work is philosophical dialogue, while the preponderance of Aristotle's is not. When this topic was mathematically removed, the centrality of topics 39 and 29 for both authors became much clearer, as figure 6 shows.

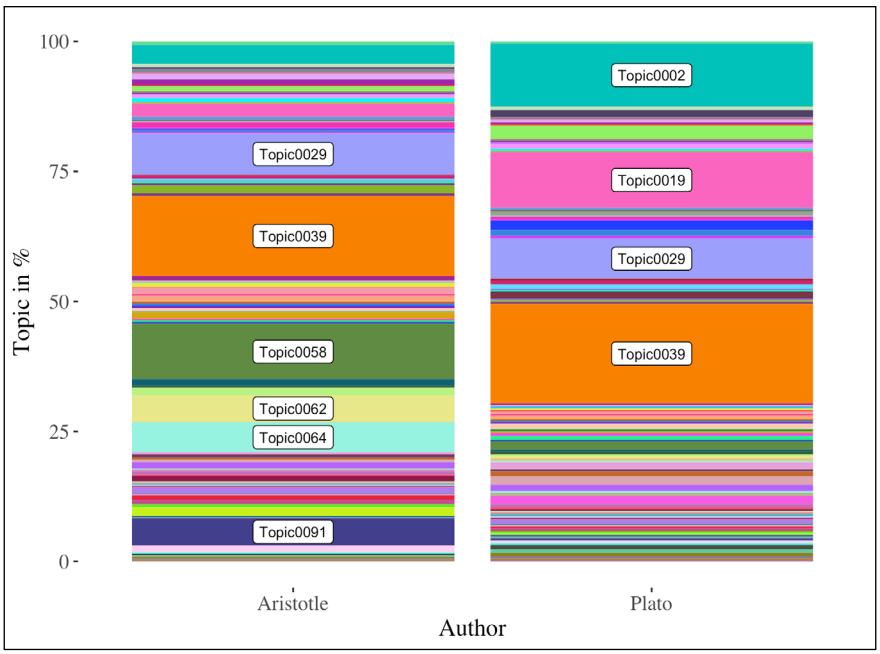

Fig. 6: Topic distribution in Aristotle and Plato after removing topic 9 mathematically.

Words strongly associated with topic 39, such as ἀρετῆς, ἀγαθὸν, φύσει, ἡδονῆς, and βίον, suggest that it is a philosophical topic dealing with virtue, good and bad, and pleasure. Words relevant to topic 29 indicate that topic deals with philosophical or scientific arguments in general. Figures 7 and 8 show the two topics, which we might also refer to as the "good and virtue" and "scientific inquiry" topics.[35]

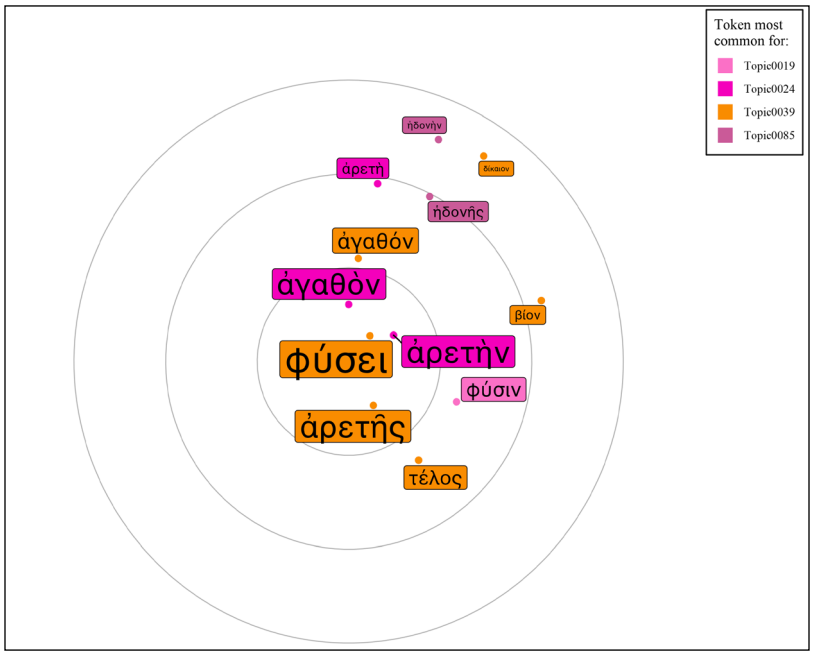

Fig. 7: Words central for topic 39, good and virtue.

35 While it is arguably more precise simply to refer to the topics by their number rather than to apply a more "limiting" label, I will here use a combination of number and short-form labels to aid readers.

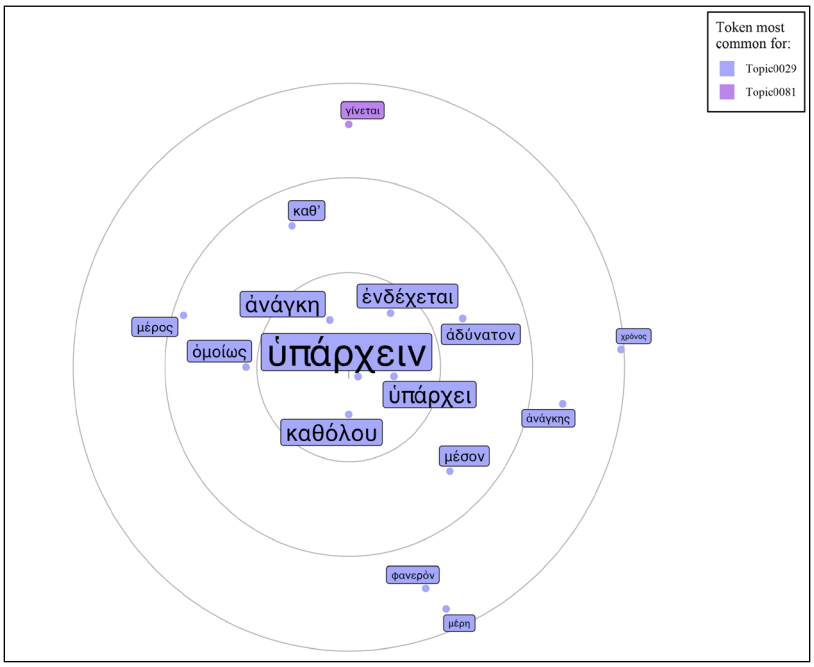

Fig. 8: Words central for topic 29, scientific inquiry.

Tracking these two topics over the whole corpus, it becomes clear that they spike within philosophical workgroups. Figure 9 shows the θ-values of the scientific inquiry (29) and good and virtue (39) topics for all passages in the First1KGreek corpus. The categories on the left are derived from the fact that the corpus itself is grouped into genre according to the bibliographic information about the workgroup. For instance, works in the workgroup *Corpus Platonicum* are considered philosophy, and their metadata reflects this. While there are obviously workgroups that feature multiple genres, in the visualisation each workgroup was only represented by the genre-label most prevalent in the workgroup. After all, one aim of this paper is to find a measurable score for passages with philosophical content, so that we might move beyond bibliographic metadata and the domain knowledge biases mentioned above.

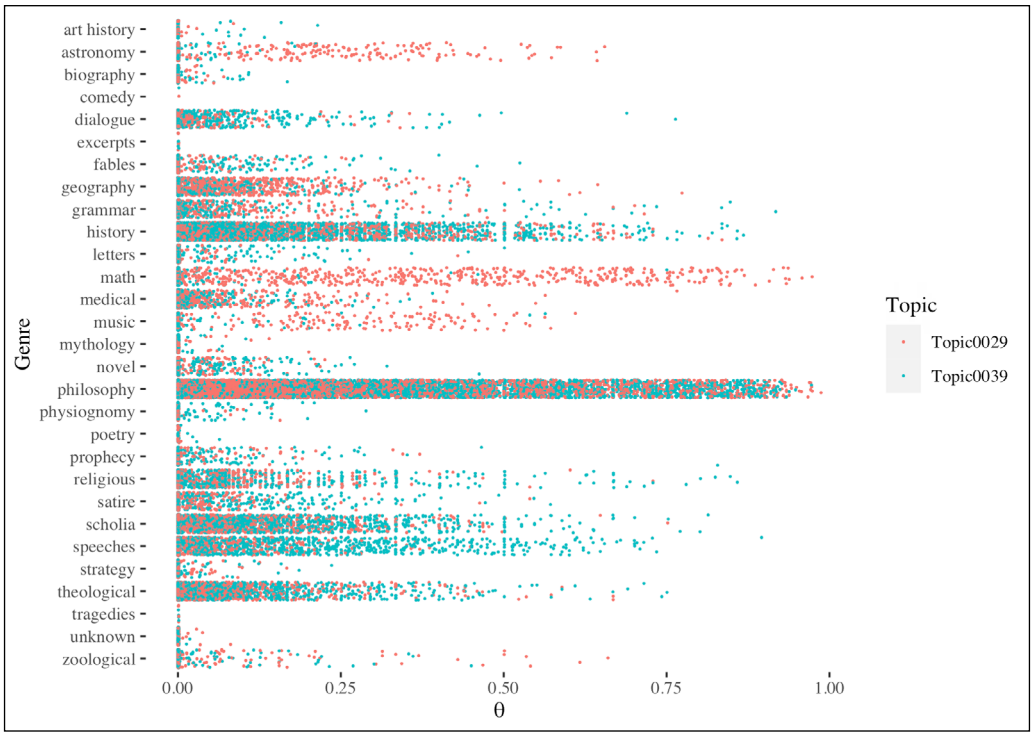

Fig. 9: Good and virtue (39) and scientific inquiry (29) topics traced over the First1KGreek corpus.

While this in itself does not prove that topics 29 and 39 suffice to generate a score that detects philosophical content, it motivated me to delve deeper. Consequently, I ran a correlation analysis between all the topics to see which topics are correlated to each other, so I could use this information to build a classification score for philosophical content using independent topics. This analysis and the observation of other topics and their correlation helped to focus on six topics: the already mentioned good and virtue (39) and scientific inquiry (29) topics, as well as topic 13, dealing with matter, substance, and subjects in a philosophical way; topic 24, dealing with good and bad; topic 47, dealing with opposites, such as black and white, and privation and possession; and finally, topic 85, dealing with pleasure and pain. Topics 29 and 30 are fully independent, while the other four topics show a mild positive correlation with one other topic each, specifically: topic 13 is mildly correlated with topic 47 and topic 24 is mildly correlated with topic 85 (see fig. 10). This makes sense when looking at the content of the topics, since the topic of "good and bad" sometimes occurs in passages where "pleasure and pain" are discussed.

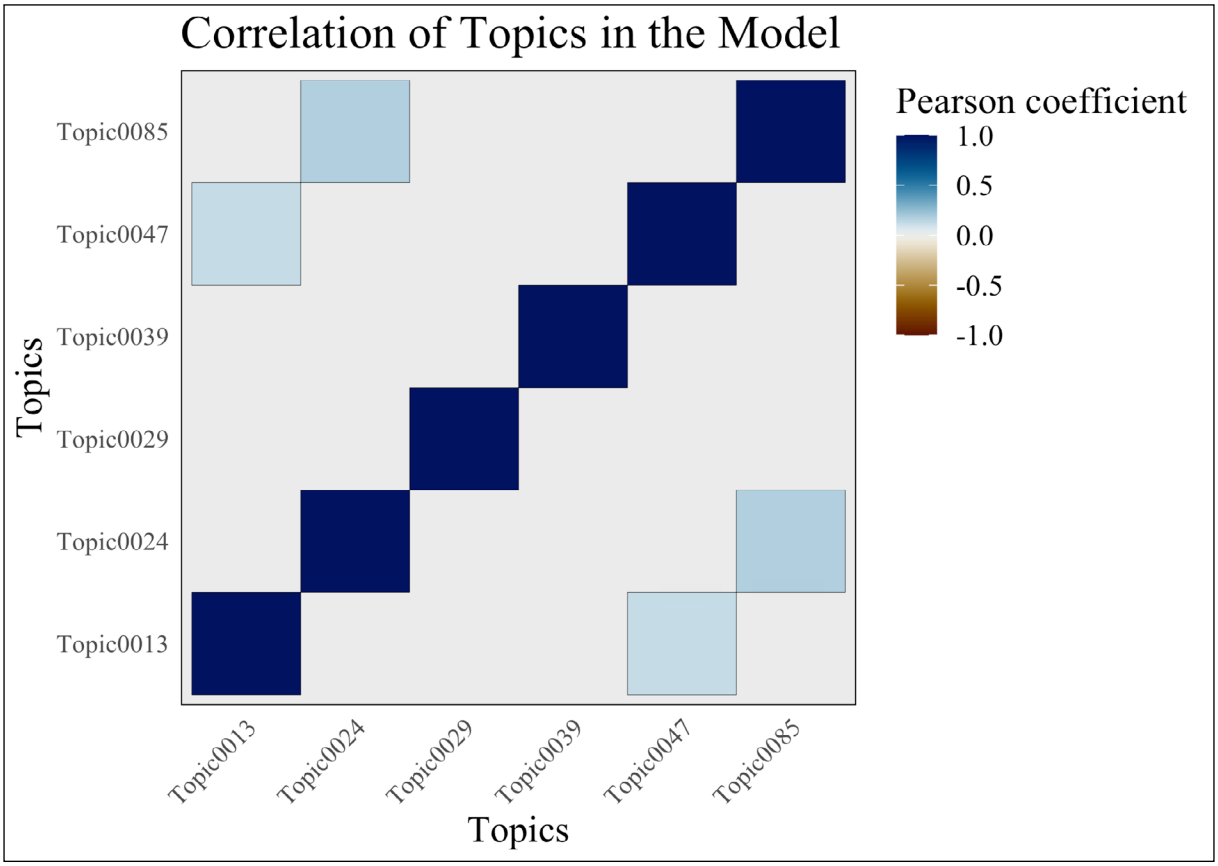

Fig. 10: Correlation of topics in the model. A Pearson coefficient around 0 signifies no correlation, a coefficient of -1 signifies a strongly negative correlation, and a coefficient of 1 signals a strongly positive correlation.

Using those six topics, I attempted to produce a classifier that would sort workgroups into philosophical and non-philosophical text. Because of the CTS/CITE framework and metadata in the Perseus Catalog (a bibliographic metadata aggregation tool that includes the Perseus Digital Library, OGL texts, and more), this is relatively simple to do.[36] That is, I compared the result of the classifier score with the qualitative judgement (contained in the metadata) as to whether a workgroup can be considered philosophical. This, in effect, meant comparing my classifier score with the assessment of philologists and then checking the precision, recall, and F1 score of that classification. The precision is the percentage of true positives among the true and false positives. The recall is the percentage of true positives among

36 See https://catalog.perseus.org [Last access 10-06-2020].

true positives and false negatives. The F1 score is the harmonic mean of recall and precision. While in data science it is important to increase the F1 score, the classifier with the greatest F1 score is not always the most useful, because it is possible that the score is over-trained for specific features in the corpus, which in our case means a very special philosophical signal works for most philosophical works, but excludes minority features. Data science is in itself experimental and accepts whichever score works best in practice.[37]

I tested three kinds of classifiers that were all based on the θ-scores of one or more of the six topics mentioned above. For two, I normalized the workgroup θ-scores by dividing through the arithmetic mean of the θ-scores in the corpus, and in one score I measured the distance of a pairwise combination of those scores to the origin. Put more accessibly, a classifier either uses the factor with which one topic or a group of topics from the identified set is more or less likely to occur in a specific workgroup in comparison to the overall corpus, or it looks at the distribution of a topic pair within a workgroup.

The first kind of classifiers simply look at whether the normalised mean of a topic within a workgroup is greater than 1. The calculated number refers to how much more or less likely a topic is to occur in any given workgroup compared to the overall corpus. This number is always positive, whereby a score of exactly 1 would mean that the topic occurs within that workgroup as often as in the overall corpus; a score between 0 and 1 shows that topic occurs less frequently in the workgroup than in the overall corpus (e.g., a score of 0.5 would mean that the topic only occurs half as often as expected given its distribution over the whole corpus); and a score greater than 1 shows the factor by which a topic occurs more often within a workgroup than in the overall corpus (e.g., a score of 2 would mean that the topic occurs twice as often).

The second kind combines several pairs of topic scores through polarization and tests whether the resulting radius is greater than 1. Put more simply, it checks whether two topics are notably prevalent in a workgroup (e.g., a pair of θ-values of 0.6 and 0.8). Polarisation is a simple way of reducing multidimensional components to two scores: the radius, which is the distance of a point to the origin, and the angle of a vector from the origin to the point. Since all θ scores are positive, the angle is here less important than the radius and was not used for this kind of classifier in this study.

The classifiers of the third kind combine the θ-scores' logarithmic normalization and comparison of pairwise maxima of the topic scores. That is, they check whether a combination of two scores is more prevalent in a workgroup than in other workgroups. These classifiers are thus an extension of the first kind of classifiers listed above, but they include more than just one topic score. For these classifiers, topic pairs were selected based on the content of a topic using Classics domain knowledge. For instance, as shown above, topics 24 and 39 seem to have clearly ethical content, while topics 13 and 29 seem to be more structural. The maxima check selects whichever value is greater in each pair, while the logarithmic normalization is just a mathematical trick to express those scores as positive or negative values instead of values that are smaller or greater than 1. This means that these classifiers measure whether either both pairs of topics are slightly or one pair extremely more prevalent in a workgroup than expected, given the topic pairs' occurrence in the overall corpus.

Table 1 shows the six highest-ranked classifiers: two are of the first kind, (1) and (4); three are of the second kind, (2), (5), and (6); and one is of the third kind (3), as explained above. Looking at the table, it becomes clear that one classifier—labelled in the table as (1)—of the first kind of classifiers actually resulted in the best F1 score. That is, when the classification was based just on topic 47 (opposites), thereby judging a text on whether it has a higher or lower than average expression of that topic, it could

37 Witten et al. (2011), 403.

achieve .93 precision (i.e., 93% of the workgroups classified as philosophy are, according to their metadata, philosophy, and thereby true positives), .71 recall (i.e., 71% of the texts that are philosophy according to their metadata have been classified as philosophy by the score), and an F1 score of .81. Topic 47 is also part of the next highest-ranked classifier (2): a polarized combination of topics 47 and 13 (matter/substance). Unfortunately, topic 47, which is the topic dealing with opposites (such as black and white and privation and possession), is much more present in Aristotle and commentaries on Aristotle than it is in the Platonic dialogues and, consequently, the classifiers that are based on topic 47 do not classify the *Corpus Platonicum* as philosophy. This classifier would be a hard sell to classicists!

	Condition	Precision	Recall	F1 Score
(1)	$\frac{\overline{\theta}_{t47}}{\mu_{t47}} > 1$	0.93	0.71	0.81
(2)	$\sqrt{\overline{\theta}^2_{t13} + \overline{\theta}^2_{t47}} > 1$	0.84	0.77	0.81
(3)	$max\left(log_2(\frac{\overline{\theta}_{t29}}{\mu_{t29}}), log_2(\frac{\overline{\theta}_{t13}}{\mu_{t13}})\right) + max\left(log_2(\frac{\overline{\theta}_{t39}}{\mu_{t39}}), log_2(\frac{\overline{\theta}_{t24}}{\mu_{t24}})\right) > 0$	0.57	0.97	0.72
(4)	$\frac{\overline{\theta}_{t13}}{\mu_{t13}} > 1$	0.95	0.57	0.71
(5)	$\sqrt{\overline{\theta}^2_{t47} + \overline{\theta}^2_{t9}} > 1$	0.61	0.86	0.71
(6)	$\sqrt{\overline{\theta}^2_{t13} + \overline{\theta}^2_{t47} + \overline{\theta}^2_{t24} + \overline{\theta}^2_{t85}} > 1$	0.49	0.83	0.62

Tab. 1: The six highest-ranked classifiers, ranked by their F1 score.

A classifier of the third kind (3), however, seems more suitable to determine whether a text is philosophy or not, despite its lower F1 score of .72. This classifier includes the topics on matter and substance (13), scientific inquiry (29), good and bad (24), and virtue (39); that is, two topics relating to scientific inquiry and two relating to good and virtue. Here the recall is significantly better, which means that most of the philosophical workgroups were correctly recognised as philosophical, but the precision is lower, which means more workgroups whose metadata does not deem them philosophical are classified as philosophical. The low precision explains why this has a lower F1 score than the score solely relying on topic 47 (opposites).

Taking the three highest-ranked classifiers (1–3) into account, as a data scientist it is tempting to simply select the one with the higher F1 score, but I am also a classicist, and I thus prefer the more complex classifier that includes four topics for the following reason. As outlined above, topic 47 includes words like "black", "white", and "opposites", which seems to be a clear Aristotelian topic; however, I also know that a lot of works in the First1KGreek corpus have been clearly influenced by Aristotle. Thus, our data is biased in favour of Aristotelian philosophy and the higher F1 score of the classifier based solely on topic 47 simply emphasises this bias. It is "over-trained" for commentaries on Aristotle and therefore the other more complex classifier should be used for further analysis.

6. Selecting and Applying the Philosophy Score

Having now found our philosophicalness score, in what follows I will simply refer to the "PhilScore", which is expressed in the left side of the equation for classifier (3) in table 1. As can be seen, the PhilScore employs the θ-scores of four topics: two good and virtue scores and two scientific inquiry scores.

One advantage of this is that it allowed me to plot all authors in the corpus in a two-dimensional coordinate system, where the axes signify the two components of the score (fig. 11). A higher position on the y-axis represents a higher "good and virtue" score, and a higher value on the x-axis represents a higher "scientific inquiry" score. The graphical representation of our condition (i.e., the sum of the two scores is bigger than 0) in table 1 is a 45-degree line through this coordinate system. The line represents the area of the graph where the sum of the good and virtue component and the scientific inquiry component of the PhilScore equals 0: according to our classifier, everything above that line is philosophy and everything below it is not. In order to see whether it functions, I differentiated the works labelled as philosophy in their metadata with red dots and those not labelled as such are shown as blue triangles. This worked well: there is only one work (pseudo-Aristotle) that lies below the line. However, there are lots of dots where the metadata says they are not by philosophical authors, but they still sit above the line threshold.

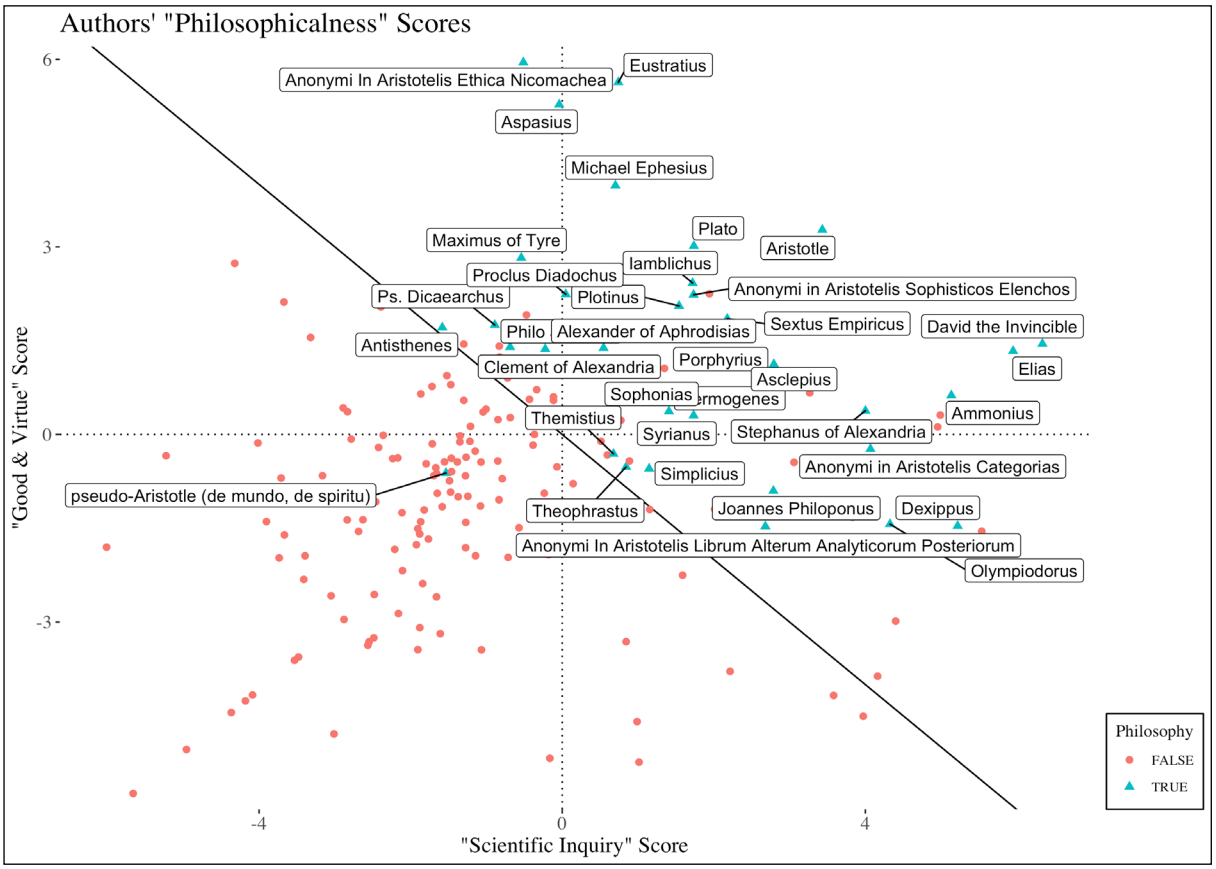

Fig. 11: Coordinate system showing the two elements of the PhilScore: the y-axis shows the good and virtue score, while the x-axis shows the scientific inquiry score. Workgroups labelled as philosophy according to their metadata are plotted as blue triangles and those that are not are plotted as red circles.

I therefore chose to look more closely at the non-philosophical works that sit above the line in order to see what they consist of (see fig. 12) according to the genre assigned in their metadata. When I looked at the points in the figure that sit above the 45-degree line (i.e., a graphical representation of our condition in table 1) and higher on the y-axis, I saw grammar, speeches, scholia, history, religious texts, and fables: that is, things that deal with good and virtue, or in the case of grammar, reference passages that deal with good and virtue when providing examples for grammatical phenomena. When I looked at the points sitting above the 45-degree line and further out to the right of the x-axis, then I saw maths, music, medical texts, and astronomy: that is, things that are highly mathematical and have to do with scientific inquiry. Thus, while these texts, according to their metadata, are not part of a workgroup that is usually associated with a philosopher, those texts still might contain philosophical content.

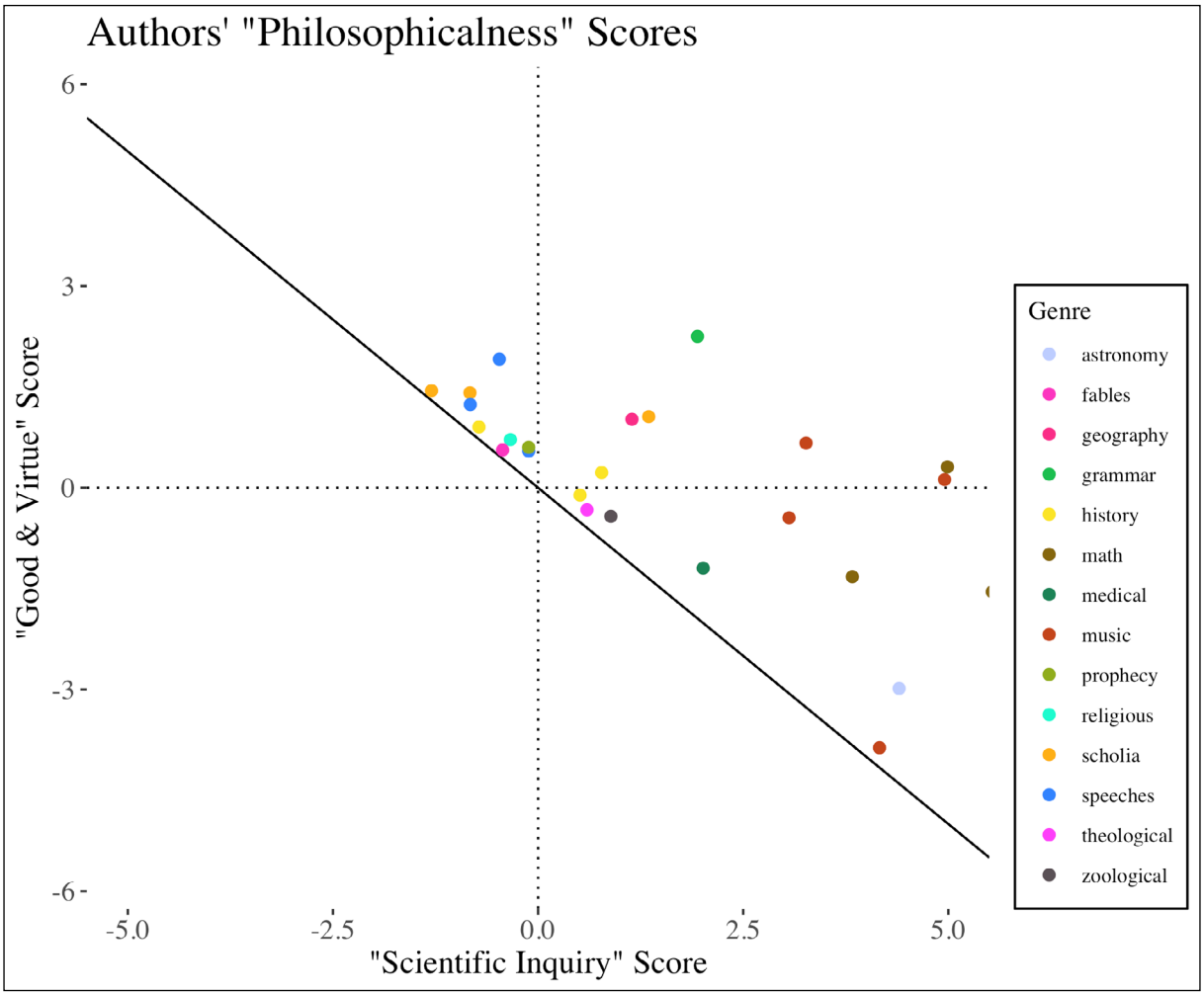

Fig. 12: Workgroups with a Philosophy Score > 0, which are not shown as philosophical based on the workgroup's metadata.

That said, it is not particularly surprising that mathematical and theological texts contain philosophical elements, and one could even argue that this could have been shown solely by bibliographic metadata, in which case topic modelling hasn't yet shown us anything new. However, because of the granularity of the First1KGreek corpus, we can shift the focus from workgroups to individual passages. After all, the topic modelling itself was performed on passages and the θ-scores for the workgroups are simply a projection. When looking at figure 13, which shows the distribution of the PhilScore for the individual passages, we see the score working. That is, the genre "philosophy" (in purple) has the highest third quartile, range, and median of the PhilScore. Thus, the genre is generally associated with a PhilScore > 0. The only four genres whose median is higher than 0 are scholia, music, zoological works, and—as the second highest—math. These are all genres very tightly connected to philosophy in the Ancient Greek canon.

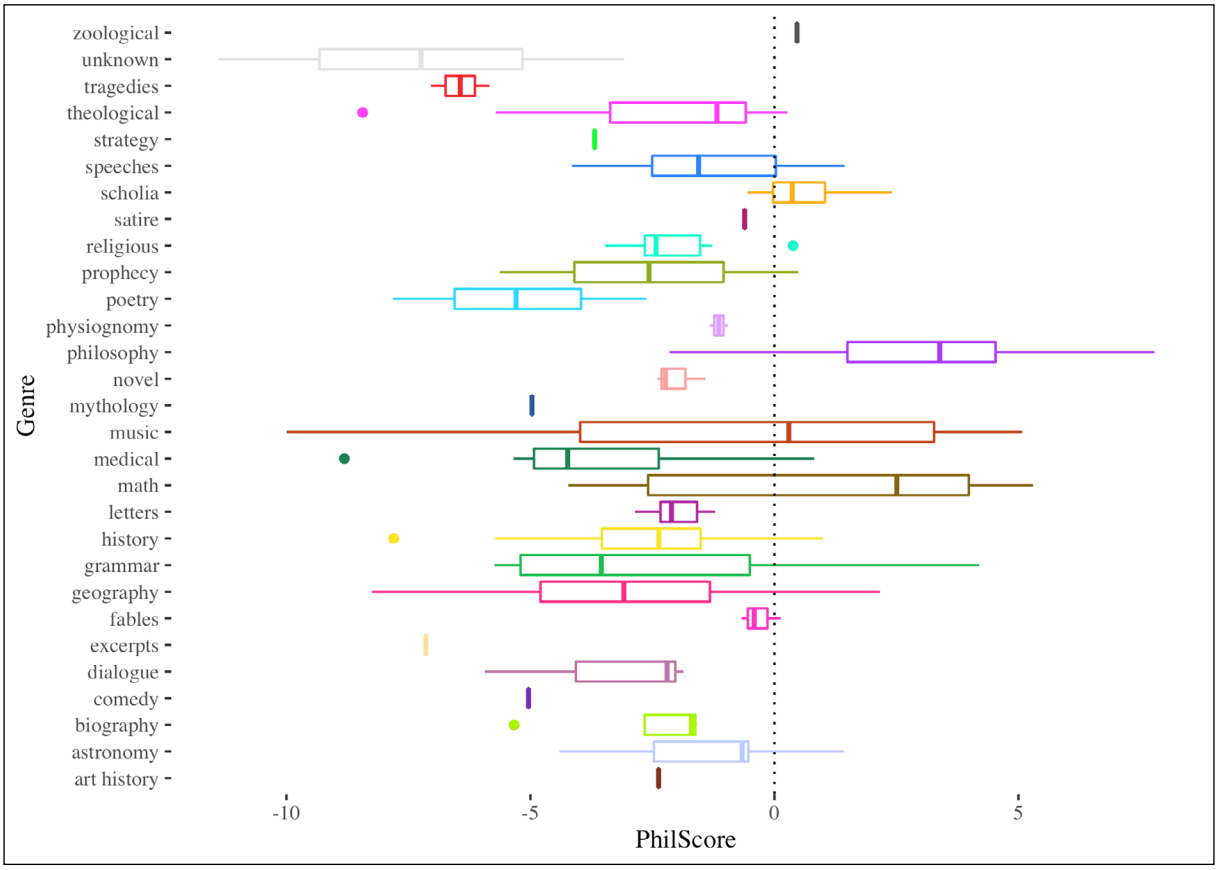

Fig. 13: Box&Whiskers-plot showing the distribution of the PhilScore over the First1KGreek corpus, based on their genre as recorded in the bibliographic metadata. The vertical bar shows the median of the distribution, the boxes refer to the first and third quartiles of the distribution, while the whiskers show the expected range of the PhilScore. Individual outliers are represented as points.

Moreover, instead of looking at a distribution boxplot, we can also look at the individual score of each passage, as shown in figure 14. Every point to the right of the dotted line (which is a graphical representation of PhilScore = 0) might contain philosophical content. This means that I can look at all the passages in all the philosophical texts that might contain philosophical content (as I could using the metadata), but more importantly, I can also look at all the passages in non-philosophical texts, that nonetheless most likely contain philosophical thinking.

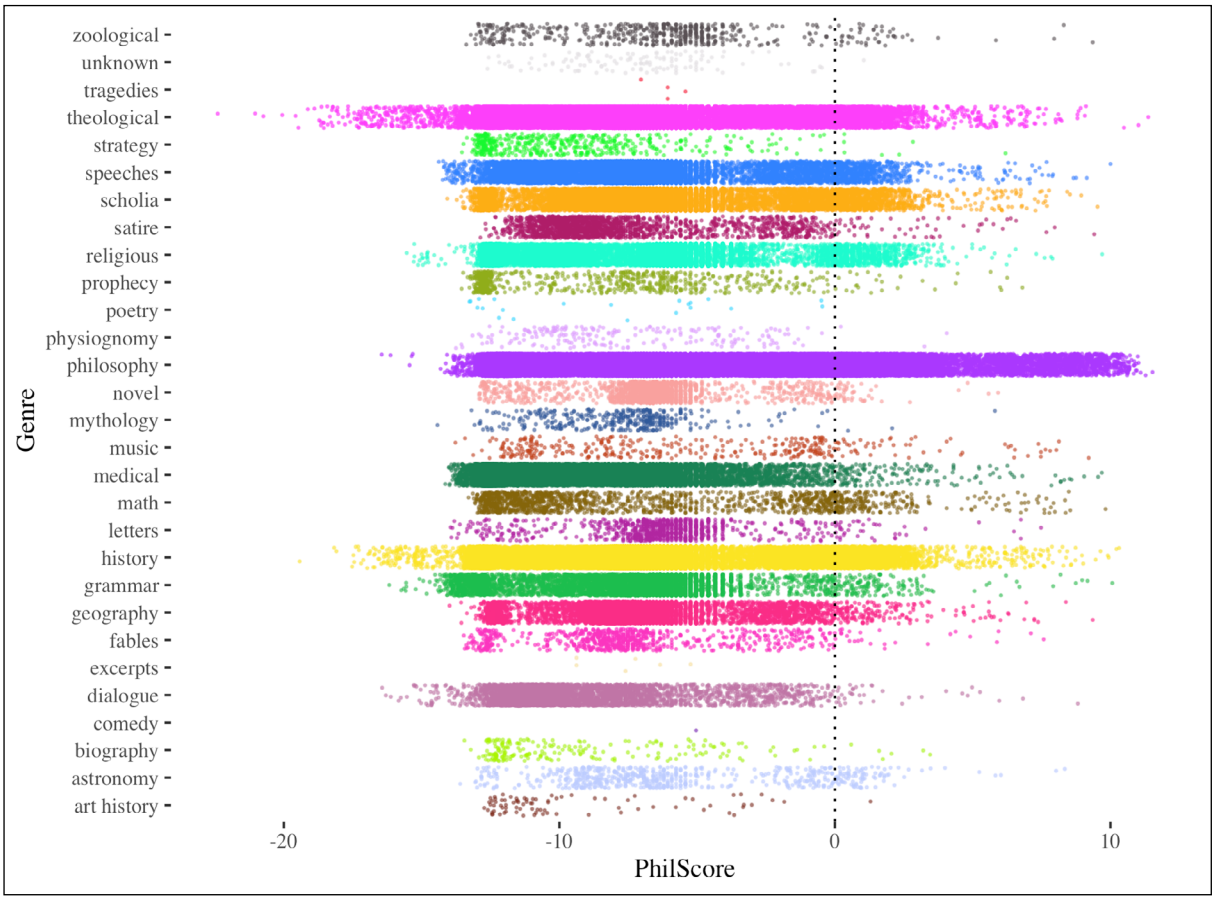

Fig. 14: Point chart showing the PhilScore of individual passages sorted by their genre as recorded in the bibliographic metadata.

Similarly, we could also apply the PhilScore not just to workgroups and individual passages, but also to the works themselves. Figure 15 shows a graphical representation of the PhilScore within the *Corpus Platonicum*, in which the preponderance is classified as philosophical, although others are clearly not. There are good philological reasons why some of the works have relatively low scores. For instance, the work with the lowest scientific inquiry score is the *Menexenus*, a work that is very unusual and has most likely not been written by Plato.[38] The authenticity of the work that has the lowest good and virtue score, Plato's *Theages*, has also been contested,[39] while the work with the highest scientific inquiry score, the *Parmenides* is considered arguably the most complex and enigmatic of the *Corpus Platonicum*.[40]

38 See Koentges (2020d) for an extensive stylometric analysis of the *Corpus Platonicum*.

39 See Kraut (1992).

40 Miller (2005).

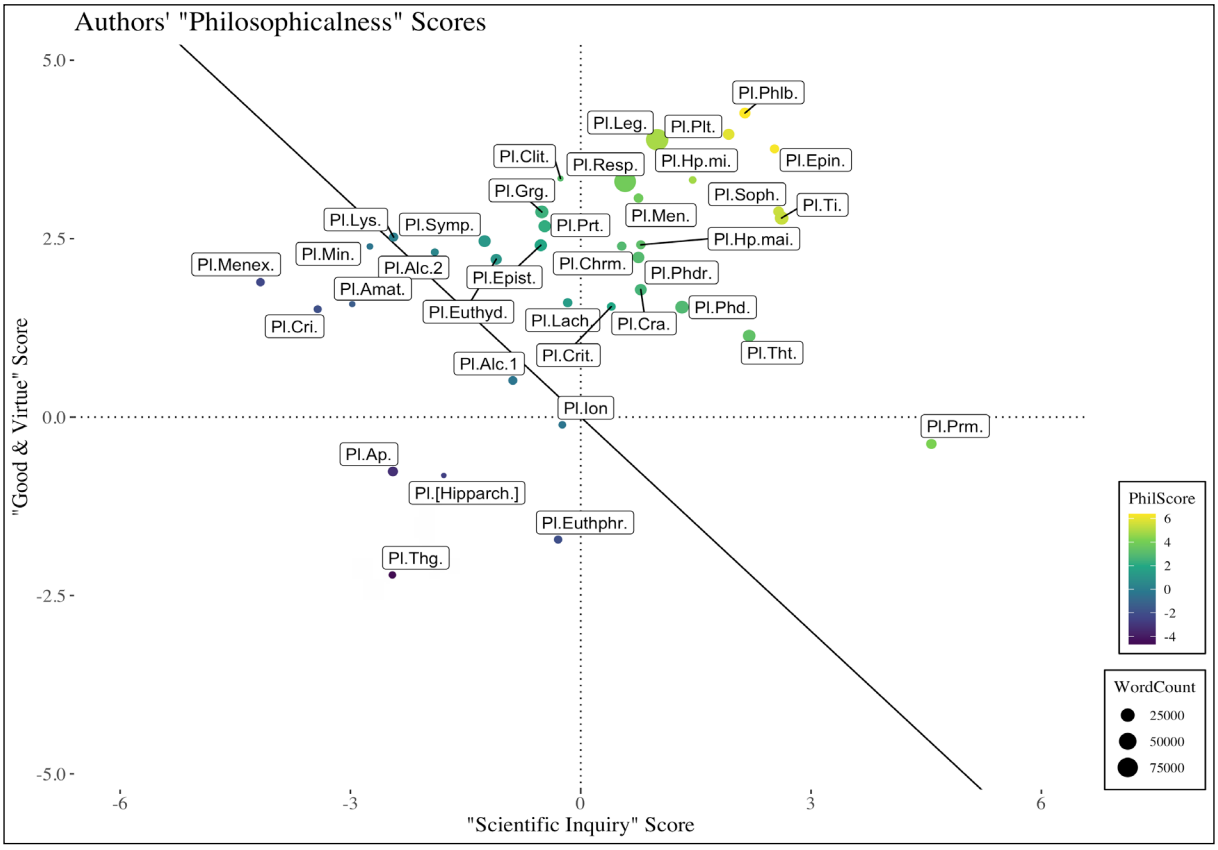

Fig. 15: PhilScore of the individual works in the *Corpus Platonicum*.

As a final point, when analysing and building the PhilScore I have only utilised the θ-scores of our topic model. However, the model also produces a corresponding topic-word-distribution, the φ-scores of the model, which I have so far only used to plot the content of a topic (figs. 5, 7, and 8). If we were to apply those φ-scores to Ancient Greek text that was not included in the First1KGreek corpus, we could also use the PhilScore to trace philosophical content extant in those texts. This means that the PhilScore, which is technically relative to the model and the texts used in this paper, could potentially be generalised for all Ancient Greek literature. That, however, should be tested in follow-up research.

7. Conclusion

In this paper I have discussed how topic modelling can be leveraged to find ways into huge language corpora that would otherwise be far too large to read and, more specifically, I have demonstrated that it is a viable way to find philosophical passages in works not written by philosophers. In order to find granular, measurable scores for genre attribution, I had to employ knowledge from different academic fields, and so this example builds on four domains: metadata generation in the Library and Information Sciences, on Data Science, historical language processing, and on Classics knowledge. Moreover, it builds on open data produced by the OGL project, without which this research would not be possible.

I showed how we can, through analysis and interpretation of a specific LDA topic model, distil a model-specific score that traces passages with philosophical content. I have also demonstrated that we cannot simply select the classifier with the highest F1 score; rather, we must employ domain knowledge and data-bias awareness to select a meaningful score. The granularity of the research corpus and the potential application of the distilled classifier to unknown text make the developed score incredibly useful for corpora too vast to read.

Despite such early, promising results, this article only represents a first step in the wider application of topic modelling for genre detection for Ancient Greek. Much more work is needed. First, we would need to test how well the projection of the score works for Ancient Greek text that was not included in the research corpus to see whether we can generalise it. Second, we should research which distance measures between the document vectors of the topic model work best in Ancient Greek. Since they are probability distributions, a sensible first attempt would be to use the Jensen-Shannon divergence to find closely related passages, as employed in my tool Metallō.[41] Because the research corpus as well as the source code and data produced in this paper, along with the article itself, are published without any major restriction of access and reuse, I hope that this work is the basis for on-going research in Digital Classics.

41 Koentges (2020b).

References

Blei (2012): D. Blei, Probabilistic Topic Models, Communications of the ACM 55 (4), 77–84.

Blei et al. (2003): D. Blei / A. Y. Ng / M. I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research 3 (January, 2003), 993–1022.

Brett (2012): M. R. Brett, Topic Modeling: A Basic Introduction, Journal of Digital Humanities 2,1 (2012), http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/ [Last access 10-06-2020].

Calame (1974): C. Calame, Réflexions sur les genres littéraires en Grèce archaïque, Quaderni Urbinati di cultura classica 17 (1974), 113–128.

Crane et al. (2020): G. R. Crane / L. Muellner / B. Robertson / A. Babeu / L. Cerrato / T. Koentges / R. Lesage / L. Stylianopoulos / J. Tauber, OpenGreekAndLatin/First1KGreek (Version 1.1.4837), data set on Zenodo, http://doi.org/10.5281/zenodo.3779102 [Last access 10-06-2020].

Dik / Whaling (2008): H. Dik / R. Whaling, Bootstrapping Classical Greek Morphology, in: [Proceedings of] Digital Humanities 2008, Oulu 2008, 105–106.

Geßner (2010): A. Geßner, Das automatische Auffinden der indirekten Überlieferung des Platonischen Timaios und die Bedeutung des Tools „Zitationsgraph“ für die Forschung, in: C. Schubert / G. Heyer (eds), Das Portal eAQUA: Neue Methoden in der geisteswissenschaftlichen Forschung I (Working Papers Contested Order No.1), Leipzig 2010, 26–41.

Harris (1954): Z. S. Harris, Distributional Structure, Word 10,2–3 (1954), 146–162.

Jockers (2013): M. L. Jockers, Macroanalysis: Digital Methods and Literary History, Urbana 2013.

Jolliffe (2002): I. T. Jolliffe, Principal Component Analysis, 2nd ed., Springer Series in Statistics, New York 2002.

Koentges (2016a): T. Koentges, Topic Modelling of Historical Languages in R, blog, http://www.dh.uni-leipzig.de/wo/topic-modelling-of-historical-languages-in-r/ [Last access 10-06-2020].

Koentges (2016b): T. Koentges, Researchers to Your Driving Seats: Building a Graphical User Interface for Multilingual Topic-Modelling in R with Shiny, Digital Humanities 2016: Conference Abstracts, Kraków 2016, 605–607.

Koentges (2018): T. Koentges, Research Report: Computational Analysis of the Corpus Platonicum, CHS Research Bulletin 6,1 (2018), http://www.chs-fellows.org/2018/04/30/report-corpus-platonicum/ [Last access 10-06-2020].

Koentges (2019): T. Koentges, ThomasK81/ToPan: A Topic Modelling Workbench for Historical Languages (Version v0.5.1), http://doi.org/10.5281/zenodo.596909 [Last access 10-06-2020].

Koentges (2020a): T. Koentges, ThomasK81/tidyToPan: Account of Monte Cristo (Version 1.0.0), on Zenodo, http://doi.org/10.5281/zenodo.3605354 [Last access 10-06-2020].

Koentges (2020b): T. Koentges, ThomasK81/Metallo: Isabella Bird (Version 1.0.0), on Zenodo, http://doi.org/10.5281/zenodo.3687367 [Last access 10-06-2020].

Koentges (2020c): T. Koentges, ThomasK81/TEItoCEX: Jerome (Version 1.0.0), on Zenodo, http://doi.org/10.5281/zenodo.3859588 [Last access 10-06-2020].

Koentges (2020d): T. Koentges, The Un-Platonic Menexenus: A Stylometric Analysis with More Data, Greek, Roman, and Byzantine Studies 60,2 (2020), 211–241, https://publications.page.link/UnplatonicGRBS [Last access 10-06-2020]

Koentges (2020e): T. Koentges, ThomasK81/MeasuringPhilosophyFirst1KGreek: DCO version (Version 1.0.0), on Zenodo, http://doi.org/10.5281/zenodo.3878753 [Last access 10-06-2020].

Koentges et al. (forthcoming 2020): T. Koentges / C. Blackwell / J. Tauber / N. Smith / G. R. Crane, The CITE Architecture: Q&A Regarding CTS and CITE, in: S. Bond, P. Dilley, and R. Horne (eds), Linked Open Data for the Ancient World: A Cookbook, New York 2020.

Kusner et al. (2015): M. J. Kusner / Y. Sun / N. I. Kolkin / K. Q. Weinberger, From Word Embeddings to Document Distances, Proceedings of the 32. International Conference on Machine Learning, Lille 2015, 957–966.

Kraut (1992): R. Kraut, Introduction to the Study of Plato, in: R. Kraut (ed.), The Cambridge Companion to Plato, Cambridge 1992, 1–50.

Maaten / Hinton (2008): L. van der Maaten / G. Hinton, Visualizing Data Using t-SNE, Journal of Machine Learning Research 9 (November 2008), 2579–2605.

Miller (2005): M. H. Miller, Plato's Parmenides: The Conversion of the Soul, University Park 2005.

Nagy (1994): G. Nagy, Genre and Occasion, Mètis: Anthropologie des mondes grecs anciens 9,1 (1994), 11–25.

Pöckelmann et al. (2017): M. Pöckelmann / J. Ritter / E. Wöckener-Gade / C. Schubert, Paraphrasensuche mittels word2vec und der Word Mover's Distance im Altgriechischen Digital Classics Online 3,3 (2017), 24–36. https://doi.org/10.11588/dco.2017.0.40185 [Last access 10-06-2020].

Posner (2012): M. Posner, Very Basic Strategies for Interpreting Results from the Topic Modeling Tool, blog, http://miriamposner.com/blog/very-basic-strategies-for-interpreting-results-from-the-topic-modeling-tool/ [Last access 10-06-2020].

Rotstein (2012): A. Rotstein, Mousikoi Agones and the Conceptualization of Genre in Ancient Greece, Classical Antiquity 31,1 (2012), 92–127.

Schöch (2017): C. Schöch, Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama, DHQ: Digital Humanities Quarterly 11,2 (2017), http://digitalhumanities.org:8081/dhq/vol/11/2/000291/000291.html [Last access 10-06-2020].

Schubert et al. (2019): C. Schubert / P. Molitor / J. Ritter / J. Scharloth / Kurt Sier, Platon Digital: Tradition und Rezeption, Digital Classics Books 3, Heidelberg 2019, https://doi.org/10.11588/propylaeum.451 [Last access 10-06-2020].

Sievert / Shirley (2014): C. Sievert / K. Shirley, LDAvis: A Method for Visualizing and Interpreting Topics, Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore 2014, 63–70.

Smith (2016): N. Smith, Morphological Analysis of Historical Languages, Bulletin of the Institute of Classical Studies 59,2 (2016), 89–102, https://doi.org/10.1111/j.2041-5370.2016.12040.x [Last access 10-06-2020].

Springmann et al. (2016): U. Springmann / H. Schmid / D. Najock, LatMor: A Latin Finite-State Morphology Encoding Vowel Quantity, Open Linguistics 2,1 (2016), 386–392.

Stamatatos et al. (2000): E. Stamatatos / N. Fakotakis / G. Kokkinakis, Automatic Text Categorization in Terms of Genre and Author, Computational Linguistics 26,4 (2000), 471–495.

Van Dijk (1997): G.-J. Van Dijk, Ainoi, Logoi, Mythoi: Fables in Archaic, Classical, and Hellenistic Greek Literature: With a Study of the Theory and Terminology of the Genre, Leiden 1997.

Wallach et al. (2009): H. M. Wallach / I. Murray / R. Salakhutdinov / D. Mimno, Evaluation Methods for Topic Models, Proceedings of the 26th Annual International Conference on Machine Learning, Montreal 2009, 1105–1112.

Wishart / Prokopidis (2017): R. Wishart / P. Prokopidis, Topic Modelling Experiments on Hellenistic Corpora, Proceedings of the Workshop on Corpora in the Digital Humanities: CDH 2017, 39–47.

Witten et al. (2011): I. H. Witten / E. Frank / M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed., San Fransisco 2011.

Author contact information[42]

Dr. Thomas Koentges

Fellow in Historical Language Processing and Data Analysis
Center for Hellenic Studies, Harvard University
Assistant Professor (Akademischer Rat) at the Department of Digital Humanities,
Institute for Computer Science & Mathematics, University of Leipzig

https://www.thomaskoentges.com

42 The rights pertaining to content, text, graphics, and images, unless otherwise noted, are reserved by the author. This contribution is licensed under CC BY 4.0.