

Innovation in Loops: Developing Tools and Redefining Theories within the Project 'Digital Plato' (Digital Plato)

Eva Wöckener-Gade, Marcus Pöckelmann

Abstract *One of the great desiderata in the field of Classics unfulfilled for centuries has been to trace the reception of the tremendously influential thinker and writer Plato throughout antiquity. The task is not only challenging because of the vast amount of literature (even if we focus on Greek) but has also been impossible to solve with digital approaches so far, especially because many voices of this reception neither explicitly refer to Plato nor cite his words verbatim. A precondition to further explore these instances was, therefore, to be able to detect them. In an iterative process of discussion, implementation, and evaluation, the project developed two web-based solutions for paraphrase search on the Ancient Greek corpus fit for different research questions (based on close and distant reading) together with different supportive resources and tools. The shaping of the theoretical framework, hand in hand with the work on the algorithms, did not only yield innovative digital methods for searching Ancient Greek texts but also sparked a reconsideration of the literary concepts of 'paraphrase' and 'intertextuality'.*

Project description

The Digital Plato project, led by Charlotte Schubert, Paul Molitor, Jörg Ritter, Joachim Scharloth, and Kurt Sier, aimed to explore the aftermath and reception of Plato in Ancient Greek literature through the paraphrasing tradition of Plato's oeuvre in antiquity. The influence of the outstanding thinker (428/427 – 348/347 BC), who laid the foundation for many scientific disciplines, philosophy being one of them, is not easy to trace or to grasp: his impact on Greek literature is inherent in countless texts and manifests itself in different ways. Based on existing approaches from paraphrase detection and adjacent fields, we have developed an innovative paraphrase search for Ancient Greek with the help of which not only verbatim citations but also non-literal references such as allusions can be detected.

The disciplines involved are Ancient History, Classical Philology, Computer Science, and Corpus Linguistics. The team comprised four working groups, two from

the University of Leipzig, one from the University of Halle, and one from the University of Dresden.

Our research question addressed Greek literature from the whole period of antiquity and into the Byzantine era, and we built our corpus respectively: As our database, we used a cleaned and adapted version of the corpus of Ancient Greek literature based on the Thesaurus Linguae Graecae (TLG, <https://stephanus.tlg.u-ci.edu/>, Version E), which includes texts from the period approximately between the 8th c. BC and 16th c. AD. Besides some smaller prerequisites (e.g. a lemmatiser identifying the basic form for any given grammatical form for Ancient Greek and a CTS-URN-generator for referencing text passages precisely with unique *Uniform Resource Names*), the tools developed by the project for working with this corpus encompass mainly two types of paraphrase search and the ‘Reference-Annotator’. With the paraphrase search via rWMD (relaxed Word Mover’s Distance), paraphrase candidates characterised by semantic similarity to a passage defined by the user can be identified. In this paper, we will focus on this method; see below for more information. With the second approach for finding paraphrases, which is based on complex n-grams (i.e. sequences of diverse linguistic components), longer text segments, for example works from different authors, can be searched for paraphrase candidates; here, different paraphrase terms for the respective search queries can be modelled by adjusting various parameters. With the Reference-Annotator, the relations between text passages identified as intertextual references, such as paraphrases, can be annotated, analysed, and categorised according to linguistic and literary criteria; see below for a glimpse of the tool. In our web portal, all text passages are provided with persistent identifiers called CTS-URNs, according to the protocol of Canonical Text Services¹, to ensure a stable citation method on the level of single words (instead of page/chapter/line which would not be precise enough for our means).

By implementing different approaches, we mixed methods taken not only from Computer Science and Computational Linguistics but also from hermeneutics. The paraphrase search via complex n-grams is suitable for distant reading while the rWMD-search and the Reference-Annotator are optimised for close reading. Thus, the user can search for references to a special passage from Plato in which he is interested and analyse the results in detail with the latter two tools; on the other hand, he can compare longer texts, for example, a Platonic dialogue with a later work to detect similarities with the former. Different types of visualisations offer the user an alternative and intuitive approach to the results.

1 Christopher Blackwell and David N. Smith, “The Canonical Text Services protocol”, GitHub cite-architecture, last modified Apr 10, 2014. https://github.com/cite-architecture/cts_spec/blob/master/md/specification.md.

Our tools are implemented into a freely accessible web interface specially designed for this purpose (<https://paraphrasis.org/>). The application is not limited to the domain of Plato's works, but allows searching in an author-independent manner within the corpus. Moreover, not only references but also models can be detected since the similarity of the passages compared is the crucial factor (see below). Since the TLG-E is licensed, we chose to offer a test version with texts from a free corpus (based on the Perseus Digital Library) to all users and a full version to those who hold the licence for the TLG (on request). In addition to this portal, according to the FAIR data principles,² we provide further information about our research and access to various project-specific resources on our homepage (<https://digital-plato.org/>).

The project contributed to a multitude of research questions in different areas. We found and collected previously undetected paraphrases of Plato and other authors—which was obviously one of the main aims of our project. The analysis of these paraphrases and the work with the digital tools have stimulated us to rethink both the concepts of paraphrase and intertextual phenomena in literature. Some representative findings have been published by the project in a monograph with contributions from all disciplines.³

Methodical background: how to process the literary phenomena intertextuality and paraphrase

The long-term objective of our project was to disclose the multiple forms of Platonic reception in Ancient Greek literature with the aid of digital tools. While obvious forms such as citations are easily detected both digitally and traditionally, indirect forms like allusions pose greater problems. However, all of them expose a certain degree of similarity with their model; otherwise they could not be discovered without further indication in the context. Where this similarity does not involve exact matching of words (citation), we usually find variations of the original wording (paraphrase). An important part of our methodological reflection was, therefore, concerned with the phenomenon *intertextuality* and with *paraphrase* as a specimen of it, and how (or if) they can be processed, which was crucial for our research question: Developing a substantial understanding of what we were searching for was, of course, a precondition for detecting paraphrases and similar forms of intertextual relations.

2 Mark D. Wilkinson, Michel Dumontier, IJsbrand J Aalbersberg, et al. "The FAIR Guiding Principles for scientific data management and stewardship". *Sci Data* 3, 160018 (2016): n. pag.

3 See Charlotte Schubert, Paul Molitor, Jörg Ritter, Joachim Scharloth, and Kurt Sier (ed.), *Platon Digital: Tradition und Rezeption* (Digital Classics Books, Band 3, Heidelberg: Propylaeum, 2019).

Intertextuality encompasses the various forms of relations between different texts and was first introduced by post-structuralist Julia Kristeva as a concept inherent to any text.⁴ It has fuelled numerous innovative works on possible interconnections between texts from a new perspective in the Humanities, while former scholars had focused almost exclusively on the author and his intention. However, the post-structuralist conception of intertextuality was so universal that its application was in danger of becoming arbitrary and soon scientific discussions about a possible definition arose. In their course, scholars specified criteria to enlarge the significance and the applicability of the concept of intertextuality, reintegrating structuralist elements, especially regarding the role of the author, into it. Influential until today is Pfister's methodical approach to measuring the degree of intertextuality by using six criteria.⁵ Since many of these criteria depend on the interpretation of the reader, they yield somewhat clearer but still partly biased results and are thus hardly applicable for quantitative analysis, as Pfister himself willingly admitted.⁶ Consequently, we could only use them for the interpretation of the results of our paraphrase search in Ancient Greek literature—that is for qualitative insights, but not for their generation.

In contrast to intertextuality, which is rooted in literary studies, *paraphrase* has been investigated mainly as a linguistic phenomenon in the context of Natural Language Processing (NLP). Starting from fairly general definitions as well, for example “approximate conceptual equivalence among outwardly different material”,⁷ various divergent approaches to systematisations tried to identify the different possible types of transformations involved in paraphrasing.⁸ Contrary to most of the studies on intertextuality, the categorisations and taxonomies resulting from these studies were built with the declared goal of ‘operationalisability’, that is, to be adapted for quantitative analysis. When we tried to apply these categorisations to the analysis of

4 Julia Kristeva, “Wort, Dialog und Roman bei Bachtin (1967)”, in: Jens Ihwe (ed.): *Literaturwissenschaft und Linguistik. Ergebnisse und Perspektiven. Band 3: Zur linguistischen Basis der Literaturwissenschaft II* (Frankfurt am Main: Athenäum, 1972), 345–375.

5 Manfred Pfister, “Konzepte der Intertextualität”. In *Intertextualität: Formen, Funktionen, anglistische Fallstudien*, ed. Ulrich Broich and Manfred Pfister (Tübingen: De Gruyter, 1985), 1–30.

6 Pfister 1985, 30.

7 Robert-Alain DeBeaugrande, and Wolfgang U. Dressler, “Einführung in die Textlinguistik“, *Konzepte der Sprach- und Literaturwissenschaft* 28 (Tübingen: De Gruyter, 1981), 50. For an overview of the different proposals cf. Ho et al 2012, 856 et seqq.: Chuk F. Ho, Masrah A. A. Murad, Shyamala, Doraisamy, and Rabiah A. Kadir, “Extracting lexical and phrasal paraphrases: a review of the literature”, *Artificial Intelligence Review*, 42.4 (2012), 851–894.

8 Marta Vila, M. Antonia Martí, and Horacio Rodríguez, “Is This a Paraphrase? What Kind? Paraphrase Boundaries and Typology”, *Open Journal of Modern Linguistics* 4 (2014), 205–218, here 208–209, 211–212; Rahul Bhagat, and Eduard Hovy, “What Is a Paraphrase?”, *Computational Linguistics* 39 (2013), 463–472.

our gold standard set of paraphrases, which we had built for the statistical evaluation of all our approaches, we found that many of these supposedly ‘hard’ categories still left much room for interpretation, such as, how exactly a synonym should be defined and by whom. Questions like this one are far from being trivial for scholars working with literature in a foreign language from another time and culture with no native speakers to interview. Also, answering such questions is crucial for the generation of resources needed in NLP.

Moreover, many instances of paraphrases from our (in every imaginable way heterogeneous) corpus of Ancient Greek texts still proved to be far too complex to be described within the limits of these (already quite refined) categorisations. One obstacle is posed by the fact that they were developed on the basis of and for the performance on large, rather homogenous corpora of modern English. Still, the limited applicability would probably hold true for many other corpora encompassing literary texts regardless of the language or period. Regarding both problems, a more profound exchange between the disciplines is called for in which insights from both fields, NLP and the humanities, are taken into account and discrepancies are not marginalised or glossed over to achieve applicability of existing approaches. At least for our team, a fresh look at the problem and an intensified dialogue between the disciplines were fruitful to overcome these obstacles. In the course of our work, we achieved better results by applying ‘softer’ categories not only to the analysis of paraphrases but also as the basis of the tools developed to detect them.

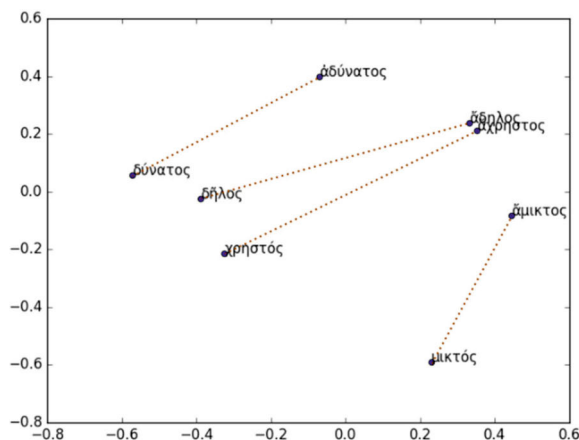
Two loops: old problems lead to new digital methods that, in turn, lead to new concepts

The incompatibility of the existing scientific approaches from both fields—literary studies and linguistics—with each other and with our research question had practical and theoretical consequences. First, we decided not to try to transfer too specific rules from other systems to ours but instead to adapt an approach rooted in very essential assumptions. This is particularly evident in our newly developed rWMD-search which is described in greater detail below. Based on the distributional hypothesis⁹ that words occurring in the same contexts tend to have similar meanings, a multidimensional vector space has been built by us representing the text corpus and all its words: a specific vector is assigned to every word in a corpus in such a way that words with similar contexts (and thus, similar meanings) are placed close to one another within this embedding. Thus, the similarity of two words can later be determined by applying a distance measure to their vectors, that is, by measuring their proximity in the embedding (cf. below). Another useful property of such

9 Zellig S. Harris, “Distributional structure”, *Word* 10 (23) (1954), 146–162.

an embedding is that it reflects not only semantic similarity but also semantic relations, as illustrated by the example in figure 1 (the vector space with originally 100 dimensions is here converted with a principal component analysis (PCA) into a two-dimensional one for the sake of presentability): it shows the relations between four Greek adjectives and their antonyms as dotted lines. Note that the lines form a pattern: it indicates that the semantic relation between adjective and antonym is similar for each pair which becomes measurable (and in this case visible) in the vector space.

Fig. 1: A convenient 2d-PCA projection of four pairs of Greek adjectives with their antonyms (from left to right): δύνατος (possible) – ἀδύνατος (impossible), δῆλος (clear) – ἄδηλος (unclear), χρηστός (useful) – ἄχρηστος (useless), μικτός (mixed) – ἄμικτος (unmixed).



We achieved a reliable word embedding of our corpus (the TLG), which means assigning a vector to each word depending on the contexts in which it is used, with the aid of word2vec¹⁰ in combination with a comprehensive semi-automatic evaluation of different parameter sets and normalisations. The principle of measuring the similarity of the meaning of two words by their proximity in the embedding can be

10 Thomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space", *Proceedings of the International Conference on Learning Representations in Scottsdale, Arizona* (2013), n. pag.

extended to whole phrases, for example, with the Word Mover's Distance (WMD),¹¹ which we use for our approach to paraphrase search. By applying such a measure, it is possible to automatically compare two phrases with respect to their meaning. In simplified terms, the WMD matches the individual words of the two phrases so that they form pairs (with the most similar ones being coupled in respect of their distance within the embedding); the average of the distances among these pairs represents the distance between the two phrases: the smaller the distance, the more similar the passages. For instance, the distance between a word and a verbatim citation of it naturally amounts to zero. If the word is used in another form or a synonym is applied instead, the distance is small; and if the word is exchanged for a semantically unrelated expression, the distance is large.

The tool which is optimised to work on the Ancient Greek corpus can be applied to a given passage of Plato in order to search for similar phrases within the corpus. It is also applicable to other research questions which involve searches for similar passages, as for example the search for sources of a given passus instead of adaptations.

Since the WMD is very costly to calculate and millions of calculations have to be performed in order to search the entire TLG, the first version developed by our team was too slow to offer an interactive search where the reader can spontaneously repeat the search and refine the results by adapting, for example, the length of the text section he is interested in or by changing the parameters. To address this problem, we used a simplified but still reliable heuristic, the relaxed Word Mover's Distance (rWMD), to pre-filter reasonable paraphrase-candidates first. In addition, we achieved a drastic reduction in the runtime of the approach by taking advantage of a mathematical special case that allows partial precomputation and also tolerates fuzziness in the length of paraphrase candidates.¹² The latter yielded a significant improvement regarding our research questions since paraphrases and other forms of intertextual relations, especially in literature, often differ in length from the original passage they refer to because they tend to abridge or extend it. This generally poses one of the bigger challenges for automated paraphrase detection which we were able to overcome in this way.

Together with further optimisations, the rWMD-search thus became usable in real time which significantly improved user experience and the way it is applied (cf. below). Figure 2 shows an excerpt of a hit list that is displayed after a search with

11 See Matt J. Kusner, Yu Sun, Nicolas I Kolkin, and Kilian Q. Weinberger, "From word embeddings to document distances", in: Francis Bach and David Blei (ed.): *Proceedings of the 32th International Conference on Machine Learning*, Volume 37 (Lille, France: JMLR.org, 2015), 957–966.

12 See for more details Marcus Pöckelmann, Janis Dähne, Jörg Ritter and Paul Molitor, "Fast paraphrase extraction in Ancient Greek literature", *it – Information Technology* 62, 2 (2020), 75–89.

the rWMD for potential paraphrases to a passus from Plato's *Republic* (335d3–7). In the passus, Plato establishes the influential axiom that heat does not cool and the good does not harm. The passage itself occupies the top of the list (No. 1) as a perfect match with the search terms (the distance amounts to zero). In the second place, we find an adaption by the Christian author Eusebius of Caesarea (*Praeparatio Evangelica* 4.22.8.5–9.1) with many verbatim agreements.

Fig. 2: The first four hits of a larger list of paraphrase candidates generated with the rWMD-search. The hits along with their distance and location are displayed as full text together with some context and also in the normalised form processed by the algorithm. In addition, there is the option to import a hit directly into the Reference-Annotator.

Trefferliste nach Wortähnlichkeit filtern 0 - 13 Änderungen				Οἱ γὰρ θερμότητος εἶναι ἔργον ψυχῆν ἄλλα τοῦ ἐναντίου. Ναὶ. Οὐδέ ὑπέρθετος κρύειταιν ἄλλα τοῦ ἐναντίου. Πάνυ γε. Οὐδέ δὲ τὸ ἀγαθὸν βλάπτειν ἄλλα τοῦ ἐναντίου.		Θερμότητος ἔργον ψυχῆν ἐκαστοῦ ὑπερθετος ὑπάρκειται ἐκαστοῦ ἀγαθοῦ βλάπτειν ἐκαστοῦ
Nr.	Distance	Location	Fundstelle	Treffer im original mit Kontext	Distance	Treffer normalisiert
1	0.0	5-4 B.C.	PLATO <i>Republica</i> 335_d Zeile 3-7 umctc:podtfgn05tfgn000004_22_8kq3-4_22_2_9j8f8	ὑποψία δὲ οἱ δίκαιον ἀδικῶς ἢ καὶ συλλέγοντες ἀρετῇ οἱ ἀγαθοὶ κακοῖς, ἄλλα ἀδικῶνται. Οὐ γὰρ θερμότητος εἶναι ἔργον ψυχῆν ἄλλα τοῦ ἐναντίου. Ναὶ. Οὐδέ ὑπερθετος κρύειταιν ἄλλα τοῦ ἐναντίου. Πάνυ γε. Οὐδέ δὲ τὸ ἀγαθὸν βλάπτειν ἄλλα τοῦ ἐναντίου. Φαίνεται. Ὁ δὲ γε δίκαιος ἀγαθός. Πάνυ γε. Οὐκ ἄρα τὸ ἀγαθὸν βλάπτειν ἔργον. ὦ Πλάτων!	Länge: 30	Θερμότητος ἔργον ψυχῆν ἐκαστοῦ ὑπερθετος ὑπάρκειται ἐκαστοῦ ἀγαθοῦ βλάπτειν ἐκαστοῦ
2	0.4524870770757356	A.D. 4	EUSEBIUS <i>Praeparatio evang.</i> 4.22.8 Zeile 6-1 umctc:podtfgn05tfgn000004_22_8kq3-4_22_2_9j8f8	αἱ οὖν οὐκ τὸ ἀγαθὸν βλάπτει ποτε οὐδὲ τὸ κακὸν ὠρεῖται, οὐ γὰρ θερμότητος ἔργον ψυχῆν. ὡς τὸ ψυχρὸν ἄλλα τοῦ ἐναντίου οὐδὲ ψυχρότητας τὸ θερμότερον. ἄλλα τὸ ἐναντίον οὐτως οὐδὲ τὸ κακὸν τὸ βλάπτειν. ἀκατάστατος δὲ ἔργον φύσει πάντως τὸ βλεῖν. ἔστι οὖν ἂν ἡ βλεῖν οὐκ αἰσάναι ἀποστρέφεται δὲ τοῦ.	Länge: 7	ψυχῆν ἐκαστοῦ ψυχρότητας θερμότητος ἐκαστοῦ βλάπτειν
3	0.59265078644893	A.D. 2	HERCULES <i>Fragmenta ethica</i> ... 49 Zeile 22-43 umctc:podtfgn05tfgn000004_22_8kq3-4_22_2_9j8f8	ἐπὶ τὸ ἀγαθὸν ἀποστρέφεται ἂν ἴσως ὁ Πλάτωνος. ἐστὶν γὰρ λόγος. (Her. 1p. 325f) οὐ γὰρ θερμὸν φρενὶ τὸ ψυχρὸν ἄλλα τοῦ ἐναντίου οὐδὲ ψυχρὸν τὸ θερμότερον. ἄλλα τὸ ἐναντίον οὐτως οὐκ ἀποστρέφεται τὸ κακότερον. ἄλλα τὸ κακότερον καὶ γὰρ ἀγαθὸν οὐ βλάπτει, ἀλλὰ κακότερον τὸ βλάπτειν.	Länge: 5	ψυχῆν ἐκαστοῦ ψυχρὸν θερμότητος ἐκαστοῦ
4	0.63270927577933	A.D. 3	PORPHYRIUS <i>De abstinentia</i> 2_41 Zeile 6-8 umctc:podtfgn05tfgn000004_22_8kq3-4_22_2_9j8f8	ὡς θεῶν. γὰρ βλεῖν ἀγαθὸς καταστρέφεται ἂν οὐκ τὸ ἀγαθὸν βλάπτει ποτε οὐδὲ τὸ κακὸν ὠρεῖται, οὐ γὰρ θερμὸν τὸ ψυχρὸν. ὡς φρενὶ Πλάτωνος. τὸ ψυχρὸν ἄλλα τοῦ ἐναντίου οὐδὲ τὸ κακὸν τὸ βλάπτειν. ἀκατὰ- ταντος δὲ ἔργον φύσει πάντως τὸ βλεῖν. ἔστι οὖν ἂν ἡ βλεῖν οὐκ αἰσάναι ἀποστρέφεται δὲ τοῦ.	Länge: 7	Θερμότητος φρενὶ πλάτωνος ψυχῆν ἐκαστοῦ βλάπτειν

The reflection on the digital methods also had consequences for our view on the literary phenomena: we developed a new perspective on the concepts of *paraphrase* and *intertextuality* in tune with our digital approaches, especially the representation of the corpus in a vector space.¹³ Some of our methodological insights were fuelled by the evaluation of the output of our algorithms and, in turn, mainly had the following consequences for the way we designed the representation of the results: paraphrase, as a phenomenon belonging to the field of intertextuality, is intrinsically linked to the recipient.¹⁴ In literature and in other forms of communication, a text segment only becomes a paraphrase if a recipient establishes an intertextual relation between this segment and another (the supposed original phrase); otherwise, the text simply remains a phrase. Applied to literature: if for example James Joyce had designed a passage in his *Ulysses* as an allusion to Homer's *Odyssey* by paraphrasing a passus of

13 Cf. Kurt Sier and Eva Wöckener-Gade, "Paraphrase als Ähnlichkeitsbeziehung. Ein digitaler Zugang zu einem intertextuellen Phänomen", in: Charlotte Schubert, Paul Molitor, Jörg Ritter, Joachim Scharloth, and Kurt Sier (ed.): *Platon Digital: Tradition und Rezeption* (Heidelberg: Propylaeum, 2019), 23–43.

14 Cf. in general the conception of Susane Holthuis, *Intertextualität. Aspekte einer rezeptionsorientierten Konzeption* (Tübingen: Stauffenburg Verlag, 1993).

the epic but no reader had ever linked Joyce's passage back to the Homeric one, intertextuality would not have been established, and Joyce's words would have stayed a phrase of his own design for everybody except for him. Moreover, the paths of the reception of Plato are so entangled and so many waypoints have been lost that one can rarely say with certainty whether a passage directly refers to Plato or rather echoes him indirectly, for example, through an intermediate source. Thus, we chose to consider a paraphrase basically as a phrase with the potential to become a paraphrase whenever an intertextual relation is established by someone, and the author simply as the first one who did so. This recipient-centred conception is also reflected in our approaches to paraphrase search, which we conceptualised as semi-automatic: both the paraphrase searches we developed, the one via n-grams (which we cannot address in depth in this paper) and the other via rWMD, propose a list of text passages which meet certain criteria to rank as promising paraphrase candidates (cf. below). The output can be sorted according to various criteria and thus adjusted to different research questions: either according to the similarity, or going by the dates of the texts' authors so that an assessment is possible from a historical perspective. The results can be exported as csv-files (comma-separated values) for further processing.

Which of the results of the search should be considered as paraphrases wholly depends on the judgement of the individual researcher. Experts might disagree here in many cases (as we have seen in our team and in literature on paraphrasing), thus encouraging a fruitful scholarly debate. For example, a researcher may classify a passage from a later author which resembles an aphorism coined by Plato as a sophisticatedly modified paraphrase adapted directly from his dialogues; another may regard it only as a faint resonance of the passage mediated through the later tradition. Hopefully, they will discuss their divergent assessments and seek further evidence to support their views, thus contributing to the exploration of Plato's reception in antiquity.

Detecting paraphrases: between similarity and vagueness

Concerning the necessary but insufficient characteristics of a paraphrase in relation to the initial phrase (or pre-text), which we want to find automatically, both similarity and, often neglected, vagueness play a central role in their detection:¹⁵ every paraphrase exhibits a significant similarity to the pre-text which enables the recipient to establish the intertextual relation between the two passages in the first place, as outlined above. However, this similarity can appear in many diverse forms and

15 Cf. Manfred Pinkal, *Logik und Lexikon – Die Semantik des Unbestimmten* (Berlin/New York: De Gruyter, 1985).

to different degrees (vagueness)—an almost exact match with the pre-text moves it closer to a citation and is easier to detect whereas loose verbal resemblances are often found in the context of allusions and similar literary phenomena which are more challenging for the readers and their stock of knowledge. Here, the approach based on the Word Mover's Distance yielded very promising results because it works almost without predefined criteria and can detect both stronger and weaker similarities between the passages as indicated by the measured distance.

Detecting paraphrases is like fishing: new methods

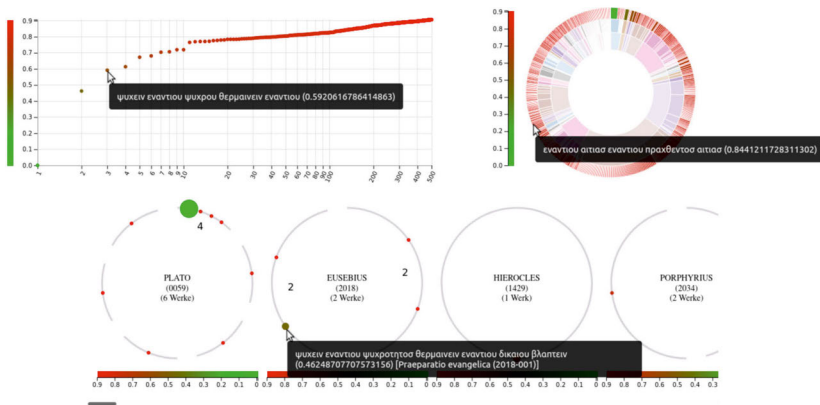
Regarding both of our approaches to paraphrase search, it is important to stress that the results they yield should never be regarded as either complete or exact which is not a weakness in our eyes but potentially advantageous. In the traditional scholarly approach, after singling out a passage or text, one would search in a very focused way for parallels in predefined texts and/or for predefined single words. In contrast, our tools use very few predefinitions. One could compare both methods to fishing with a rod (choosing a promising spot, using a specific hook and bait, devoting a lot of time, etc.) and fishing on a trawler with a net. Using the rod, with a little patience, we will probably catch the special fish we hoped for. Using the net, we catch many different types of fish at the same time: some interesting, some not, but most importantly, many unexpected ones. What was surprising for our team was the difference that the change in the speed of obtaining the results made for the way we chose and adapted our research subjects; it definitely opened new possibilities to learn by trial and error. It was also helpful for evaluating and further improving the paraphrase search and getting to know the best way to apply it since more and different types of searches (e.g. for shorter or longer phrases, using different parameters such as ignore/include stop words) could be explored and compared.

Visualising paraphrases: new perspectives

The increasing number of potential results also poses challenges to the processing of the output. Here, besides sorting the candidates suggested as paraphrases according to various parameters, visualisation comes into play. We implemented various types of visualisation to provide the researcher with a quicker and more intuitive access to the results in order to 'give him an overview'. It is important for us to clarify that also in the visualised form, the output of the paraphrase search is not conceived to represent final results, but to leave room for interpretation by the individual researcher. We tried to counteract the impression of definitiveness by implementing three types of distant reading visualisations at once which yield diverse pictures of

the transformed search results: A scatter plot for illustrating the distribution of the distance values, an interactive and multi-level sunburst diagram as an overview of the involved authors, works, and passages suggested as paraphrase-candidates, and thirdly a *planetary* visualisation as a more detailed but also more space-consuming alternative to the sunburst diagram, see figure 3 (the phrase on which the search is based is again the axiom from Plato's *Republic* 335d3–7 mentioned above).

Fig. 3: The three distant reading visualisations we integrated into our digital environment for illustrating the results of a paraphrase search. Left: a scatter plot for the distance values. Right: a sunburst diagram which aggregates authors and works for all search hits. Bottom: the search hits broken down in detail for individual authors.

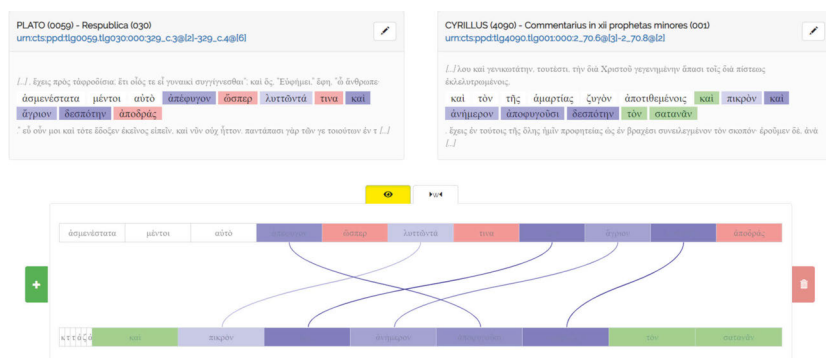


On the other hand, for the Reference-Annotator, a tool developed by us for microanalysis of confirmed paraphrases, we successfully implemented visualisations which represent the relations between a paraphrase and the original phrase in a way that makes them more easily comprehensible and comparable to other instances.¹⁶ One example is shown in figure 4: at the beginning of Plato's *Republic*, an old man relates that he is relieved to have escaped from libido like a runaway slave from a choleric and savage master (329c3–4). The literary simile has been adapted by the Church Father Cyril of Alexandria with the flight from libido being converted into the one from Satan (*Commentary on the Twelve Prophets* 70, 6–8). The Reference-Annotator allows the scholar to identify word-to-word relations between the passages and classify them according to linguistic categories (e.g. ‘same lemma, same form’ or

16 See Marcus Pöckelmann and Eva Wöckener-Gade, “Bridging the Gap between Plato and his Predecessors. Towards an Annotated Gold Standard of Intertextual References to Plato in Ancient Greek Literature”, Presentation at the EADH (2018), 07.-09.12.2018 in Galway.

'same lemma, different form', visualised in different shades of blue). Moreover, additions and omissions of words can be marked (visualised in green/red). This close-reading visualisation of the annotator helps to overcome obstacles in the analysis of paraphrases based on written text alone, which is apparent from the amount of competing linguistic approaches mentioned above.

Fig. 4: A double visualisation of the word-to-word relations between a passage from Plato's Republic and its adaption by Cyril of Alexandria in the Reference-Annotator. Above, the text of Plato is displayed on the left and the paraphrase by Cyril on the right. Different types of relations between the words can be annotated (visualised as lines in various shades of blue) and both additions (green) and omissions (red) can be marked with the tool, which results in the visualisation in the lower part of the image.



Conclusion

The project Digital Plato does illustrate how interdisciplinary teamwork and mixing methods in an iterative process can yield new insights. In our case, starting from our research question and rejecting already existing approaches as not applicable thereto, not only new and innovative digital tools were developed in accordance with our hermeneutical framework but the engagement with their theoretical foundations also led us to rethink the literary conceptions involved. The insights thus gained led to further adaptations of our algorithms and tools, and to the development of new components such as visualisations and the Reference-Annotator as a tool for the analysis of paraphrases. The insights and tools presented in this paper stem from the work of all team members, including numerous research assistants from the differ-

ent disciplines, as illustrated by our project monograph,¹⁷ who deserve our special thanks.

Bibliography

- Bhagat, Rahul, and Eduard Hovy. "What Is a Paraphrase?". *Computational Linguistics* 39 (2013): 463–472. DOI: https://doi.org/10.1162/COLI_a_00166.
- Blackwell, Christopher, and David Neel Smith, "The Canonical Text Services protocol". GitHub cite-architecture. Last modified Apr 10, 2014. https://github.com/cite-architecture/cts_spec/blob/master/md/specification.md
- De Beaugrande, Robert-Alain and Wolfgang U. Dressler. *Einführung in die Textlinguistik* (Konzepte der Sprach- und Literaturwissenschaft 28). Tübingen: De Gruyter, 1981.
- Harris, Zellig S. „Distributional structure“. *Word* 10 (23) (1954): 146–162.
- Ho, Chuk F., Masrah Azrifah Azmi Murad, Shyamala Doraisamy, and Rabiah A. Kadir. "Extracting lexical and phrasal paraphrases: a review of the literature". *Artificial Intelligence Review*, 42.4 (2012) : 851–894. DOI : <https://doi.org/10.1007/s10462-012-9357-8>.
- Holthuis, Susanne. *Intertextualität. Aspekte einer rezeptionsorientierten Konzeption*. Tübingen: Stauffenburg Verlag, 1993.
- Kristeva, Julia. "Wort, Dialog und Roman bei Bachtin (1967)". In *Literaturwissenschaft und Linguistik. Ergebnisse und Perspektiven. Band 3: Zur linguistischen Basis der Literaturwissenschaft II*, edited by Jens Ihwe, 345–375. Frankfurt am Main: Athenäum, 1972.
- Kusner, Matt J., Yu Sun, Nicholas Kolkin, and Kilian Weinberger. "From word embeddings to document distances". In *Proceedings of the 32th International Conference on Machine Learning* 37, edited by Francis Bach and David Blei, 957–966. Lille, France: JMLR.org, 2015.
- Mikolov, Thomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space". In *Proceedings of the International Conference on Learning Representations in Scottsdale, Arizona* (2013): n. pag.
- Pfister, Manfred. "Konzepte der Intertextualität". In *Intertextualität: Formen, Funktionen, anglistische Fallstudien* edited by Ulrich Broich and Manfred Pfister, 1–30. Tübingen: De Gruyter, 1985.
- Pinkal, Manfred. *Logik und Lexikon – Die Semantik des Unbestimmten*. Berlin/New York: De Gruyter, 1985.
- Pöckelmann, Marcus and Eva Wöckener-Gade. „Bridging the Gap between Plato and his Predecessors. Towards an Annotated Gold Standard of Intertextual Refer-

17 Charlotte Schubert, Paul Molitor, Jörg Ritter, Joachim Scharloth, and Kurt Sier (eds.). *Platon Digital: Tradition und Rezeption* (Heidelberg: Propylaeum, 2019).

- ences to Plato in Ancient Greek Literature". Presentation at the EADH (2018), 07.-09.12.2018 in Galway. Abstract: https://eadh2018.exordo.com/files/papers/49/final_draft/Poeckelmann_Woeckener-Gade_Bridging_the_gap.pdf
- Pöckelmann, Marcus, Janis Dähne, Jörg Ritter, and Paul Molitor. "Fast paraphrase extraction in Ancient Greek literature". *it – Information Technology* 62, 2 (2020): 75–89. DOI: <https://doi.org/10.1515/itit-2019-0042>.
- Schubert, Charlotte, Paul Molitor, Jörg Ritter, Joachim Scharloth, and Kurt Sier (eds.). *Platon Digital: Tradition und Rezeption* (Digital Classics Books, Band 3), Heidelberg: Propylaeum, 2019. <https://doi.org/10.11588/propylaeum.451>
- Sier, Kurt and Wöckener-Gade, Eva: 'Paraphrase als Ähnlichkeitsbeziehung. Ein digitaler Zugang zu einem intertextuellen Phänomen'. In: *Platon Digital: Tradition und Rezeption*, edited by Charlotte Schubert, Paul Molitor, Jörg Ritter, Joachim Scharloth, and Kurt Sier, 23–43. Heidelberg: Propylaeum, 2019. DOI: <https://doi.org/10.11588/propylaeum.451>
- Vila, Marta, M. Antonia Martí, and Horacio Rodríguez. "Is This a Paraphrase? What Kind? Paraphrase Boundaries and Typology". *Open Journal of Modern Linguistics* 4 (2014): 205–218. DOI: <https://doi.org/10.4236/ojml.2014.41016>
- Wilkinson, Mark D., Michel Dumontier, IJsbrand J. Aalbersberg et al. "The FAIR Guiding Principles for scientific data management and stewardship". *Sci Data* 3, 160018 (2016): n. pag. DOI: <https://doi.org/10.1038/sdata.2016.18>