Summary: A significant proportion of extant Ancient and Medieval Greek philosophical texts are either closely modeled on, or strongly influenced by, earlier works. Identifying which texts derive from which sources is highly valuable, as it provides deeper insight into the context, scope, and purpose of each work. Traditionally, establishing such relationships requires painstaking and timeconsuming manual comparison—an approach that often falls short of its goal. The aim of the diploma thesis is to develop a system capable of analyzing, processing, storing, and cross-referencing these texts, while also estimating potential cases of textual dependence using a range of NLP methods. These include strict token-level comparison using a BM25 model; lemma-based similarity using lemmas generated by GreCy—a Python machine-learning module trained on Ancient Greek to produce highquality lemmatizations—and subsequent BM25 computation on those lemmas; and sentence-, word-, and subword-level embeddings generated with a specially trained BERT model. The system also incorporates a novel method employing set-based models and graph structures to cross-validate results and compare standard techniques with heuristic approaches in search of an optimal balance. In addition, part-of-speechbased comparisons will be used to capture stylistic similarities. For any input text, the system will produce a ranked list of candidate source texts, each accompanied by a weighted similarity score. Although the final results may contain occasional inaccuracies, the system is designed to offer a well-informed and meaningful estimate of textual relationships. A parallel goal is to significantly reduce the burden of manual cross-referencing: instead of comparing the input text against thousands—or even tens of thousands—of potential sources, the researcher will need to examine only a short list of highly probable matches.

Textual similarity is assessed at the lexical, phrasal, and syntactic levels with corresponding weight to the final estimation:

(i) Lexical similarity

- absolute match: a word occurring in the same grammatical form (1)
- match of a word occurring in different forms (e.g. in another case, mood, or dialectal form) (0.9)
- full match of the lexeme: cognate words (sharing the same root) (0.8)
- partial match of the lexeme: a word occurring in simple form in one text and in compound form in the other, or compound with different components in each text (0.7)
- synonyms (0.6)
- synonyms by virtue of the rhetorical figure of litotes (0.5)
- synonyms by virtue of punctuation (e.g. a rhetorical question of partial ignorance occurring in one text that is equivalent to a negative statement occurring in the other) (0.4)
- antonyms (0.3)

(ii) Phrasal similarity

- sets of words that constitute a phrase and occur as elements of the phrase in the same order (1)
- sets of words that constitute a phrase and occur as elements of the phrase in a different order (0.9)
- periphrases synonymous with a single word (0.8)

(iii) Syntactic similarity

- similarities with respect to sentence structure (≤1)

- similarities with respect to period/clausal structure (≤1)