An Experiment on Plato's Gorgias as an Introduction to Textometry

Bénédicte Pincemin, Stéphane Marchand

[Accepted manuscript version 2021-09-01]
In Stelios Chronopoulos, Felix K. Maier, Anna Novokhatko, eds
Digital Text Analysis of Greek and Latin sources
***Classics@*, 20, 2022**
Washington, DC: Center for Hellenic Studies, Harvard University
https://classics-at.chs.harvard.edu/volume/classics20-digital-text-analysis-of-greek-and-latin-sources/
https://classics-at.chs.harvard.edu/an-experiment-on-platos-gorgias-as-an-introduction-to-textometry/

An Experiment on Plato's *Gorgias* as an Introduction to Textometry

Bénédicte Pincemin, Stéphane Marchand

1. Background: What is Textometry? How could it concern Classical Studies?

Current implementations of Digital Humanities for classical studies are often either rich digital editions, with much effort devoted to display a pleasant layout and to offer efficient navigation paths; or indexed corpora or databases with querying facilities, for which the focus is on searching functionalities and not on primary text visualization and reading. However, through the example of the textometric approach (presented in section 1.1) that is implemented in TXM open-source software (1.2), we would like to show that complex textual representation and computational analysis can be both taken into consideration into a unique digital framework (1.3). In order to show that such a method can offer great possibilities for researchers in the Classics field, we will give a more complete presentation of the methodology applied to Plato's *Gorgias* (section 2 and 3).

1.1 What is textometry?

Textometry is a computer-assisted textual analysis method that is based on word counts (or on any linguistic feature counts) and that links statistical processing to data referencing and contextual comparison. Main textometric functions are specificities (a keyword computation), correspondence analysis (a geometrical view of the corpus content), cooccurrences (lexical collocations), and KWIC (Keyword in context) concordances (for a technical presentation of these functions, see appendix, section 5).

The textometric methodology was founded in the 70s in France (Léon and Loiseau 2016). Benzécri and his students launched the *analyse des données* field and created new exploratory

multivariate statistical methods, mainly correspondence analysis and its combination with clustering (Benzécri 1973, 1981). Then, the Saint-Cloud Laboratory coined the word "lexicometry" (which would evolve later to "textometry" and "logometry", but the core procedures remain the same) and added new methods especially dedicated to quantitative lexical analysis: specificities (*spécificités*), collocations (*cooccurrences*) (Lafon 1984), phrase detection with repeated segments (*segments répétés*) (Salem 1987). At Nice University, Brunet was also a pioneer in developing and applying these new methods to French literature (Brunet 2011). An overview of the methods is gathered in Lebart, Salem, and Berry 1998. Moreover, since 1992, the JADT Conference (International conference on statistical analysis of textual data), which is led by the textometric scientific community, has been held every two years; its proceedings have been published online since 1998 on the Lexicometrica website.[1]

Compared to other text analysis methods, textometry is a balanced combination of quantitative and statistic methods on the one hand, and qualitative methods on the other hand. It occupies an intermediate position between text mining (which replaces the textual data by quantitative summaries, extractions, and visualizations) and annotation software (which enrich and investigate a detailed view of the text); in other words it achieves both a close reading and a distant reading.

Several textometric applications are currently available for research, like DtmVic, Hyperbase, Lexico, IRaMuTeQ, Trameur, TXM.[2] In this paper our example will be run with TXM.

1.2 What is TXM?

TXM is an open-source textometric software that was initiated in 2009 within the frame of the Textométrie project (2007–2010, founded by the ANR French agency), gathering four research laboratories, from Lyon, Paris, Nice, and Besançon. Its development is coordinated by Serge Heiden (Lyon, IHRIM laboratory), and the two main developers are Matthieu Decorde (Lyon, IHRIM Laboratory) and Sébastien Jacquot (Besançon, ELLIAD Laboratory) (Heiden 2010).[3]

---

[1] JADT conference archives: http://lexicometrica.univ-paris3.fr/jadt/index.htm

[2] Here are the websites for these applications: DtmVic: http://www.dtmvic.com ; Hyperbase: http://ancilla.unice.fr/hyperbase ; Hyperbase Web Edition: http://hyperbase.unice.fr ; IRaMuTeQ: http://www.iramuteq.org ; Lexico 5: http://www.lexi-co.com ; Trameur: http://www.tal.univ-paris3.fr/trameur ; TXM: see next footnote.

[3] Textometry project and TXM software website: http://textometrie.org

The aim of the Textometry project was for the new software to be able to fully manage and analyze state-of-art corpora, that is, structured and annotated corpora, like TEI corpora. Another aim was to launch an open-source development the partners could share.

TXM implements the textometric approach, so that it includes specificities, correspondence analysis, cooccurrences, and back-to-text facilities implemented as an advanced concordance view linked to a text view. This paper will show some other available functions too: progression in section 3.2, and index (word frequency lists) mainly in 3.4. Indexing and searching are operated by the CQP[4] component (Christ 1994; Evert and Hardie 2011), and statistical computing relies on R[5] components; processing can be extended or personalized with scripts. One of TXM main features is its ability to input many kinds of digital texts, from plain text to generic XML and XML-TEI. On the basis of these features, TXM is a fifth generation concordancer according to McEnerey and Hardie typology (McEnery and Hardie 2012).

TXM runs on Windows, Linux, MacOS, and can also be set as a web portal version, for collaborative work or on-line corpus publication.[6]

1.3 Three (Functions) in One (Digital Edition): Reading, Searching, Analyzing

As textometry pays attention both to complex queries and statistics, and to fine text reading, it can be a new unified framework for digital editions in a classical studies context. With a TXM solution, for example, humanists are not compelled to choose between either a digital edition, or a database, or a search engine, to study and publish their text corpora.

Actually, TXM is a *search engine*. Complex queries can be processed about words, phrases, morphosyntactic patterns (or patterns based on any other information that has been added to the corpus through tagging), also allowing for search patterns that include gaps (like, for instance, in ancient Greek μὲν … δὲ). Queries can cross information from every level, from lexical tags to text structures, which can be internal structures (inside texts, like chapters) or overall structures about groups of texts (like author or text genre).

---

[4] CQP stands for Corpus Query Processor, which is the name of the search engine component included in the IMS Open Corpus Workbench (CWB) (see http://cwb.sourceforge.net) initially developed at the University of Stuttgart (http://www.ims.uni-stuttgart.de/forschung/projekte/CorpusWorkbench).

[5] R is a freely available language and environment for statistical computing and graphics. CRAN (The Comprehensive R Archive Network, http://cran.r-project.org) is currently one of the main providers for open-source and up-to-date statistical packages.

[6] The developer environment and some of user documentation (user manual [Heiden, Decorde, and Jacquot 2018], user interface, mailing list) are available in English.

At the same time TXM is a *database* that manages text documents. Numerous metadata can be associated to each document (such as author, date of composition, text type, etc.), and selection of documents can be built based on the metadata's values. This can be illustrated by the text selection function in the Base de français médiéval instance of TXM web portal version (Figure 1),[7] which shows very detailed selection panels that are used like database query forms. Furthermore, any textual unit, from words or text chunks to text groups, can be described with metadata information, and then be used to target a selection in combination with any information on any text level. For instance, one may focus on sentences which include some linguistic pattern and which were written within a given time span.

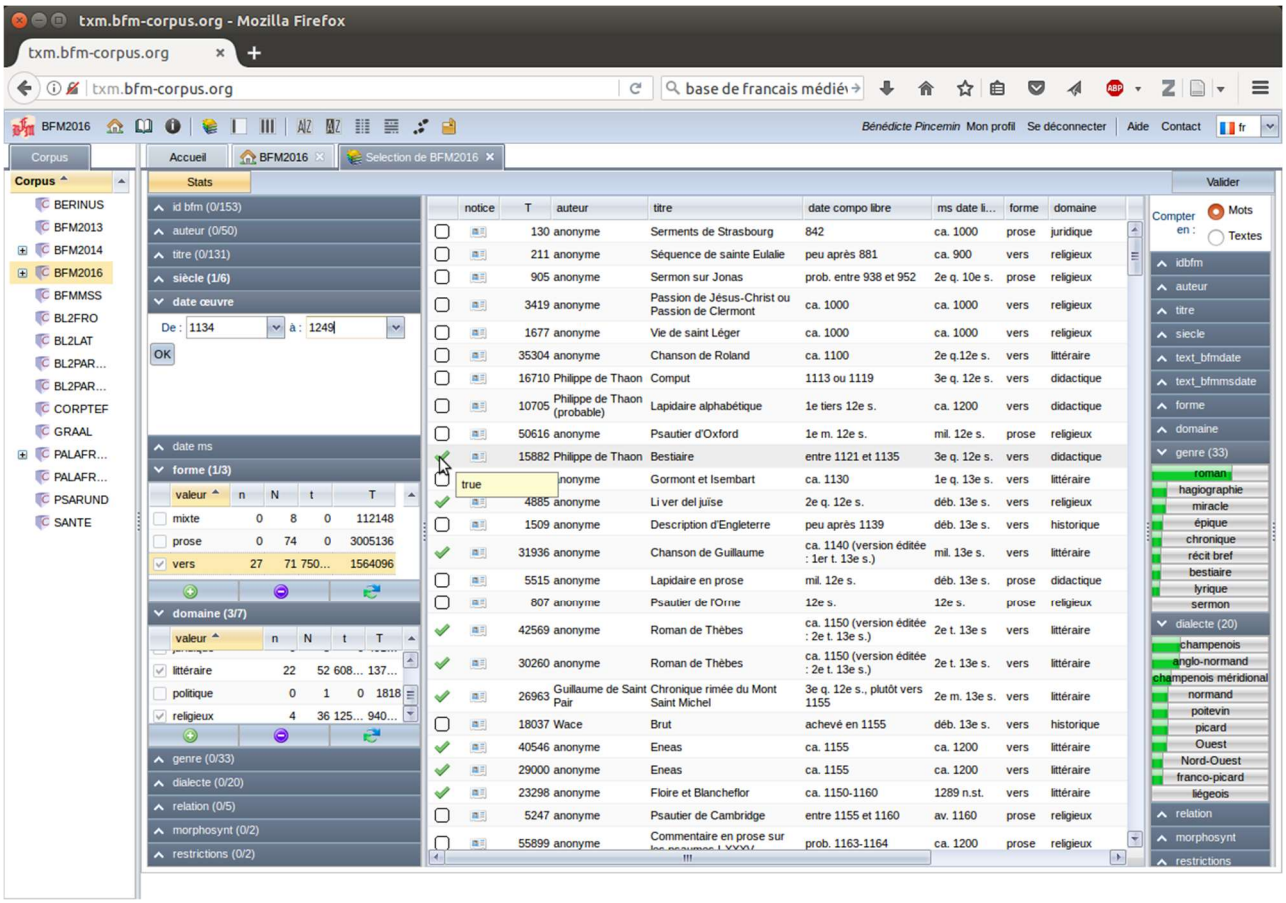


Figure 1. Text selection in the Base de Français Médiéval TXM web portal.

Then, TXM provides a wide range of corpus linguistic functions to process these lexical and textual selections (word frequency lists, KWIC concordance, collocations, keyword analysis, distributions, multidimensional analysis, etc.; see section 1.2). These analytic features can be coupled with a fine text encoding (like XML-TEI), making it possible to record precise philologic

[7] Base de français médiéval: http://bfm.ens-lyon.fr

information[8] and to build fine-tuned HTML *editions* to visualize the text (Figure 2). One important feature of the process is the possibility to manage and distinguish information used as text content (words to be searched and counted), and information useful to follow the history of the text, assess its structure, and assist text reading and interpretation (critical apparatus, editorial text divisions, speech turn labels, etc.). If available, facsimiles of source documents can be displayed so that the researcher still has a view on the primary document. Thus one keeps access to information that might have been lost because of necessary digital encoding choices.

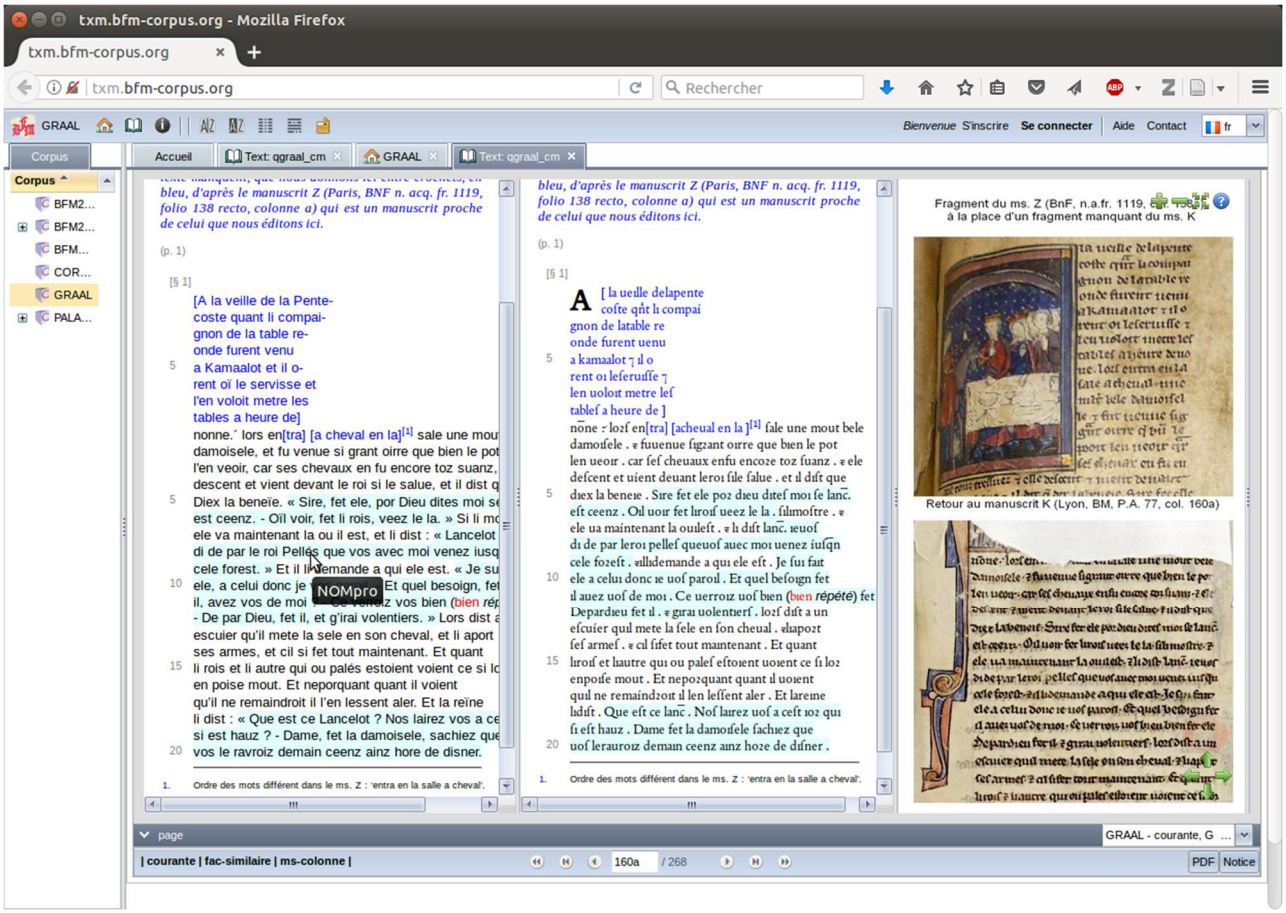


Figure 2. Synoptic views of the *Queste del Saint Graal* in a TXM web portal edition (Marchello-Nizia and Lavrentiev 2013): here two kinds of transcriptions and the manuscript image are aligned.

Thus with this example we see that it is possible to give the same text a unified digital edition that offers all together searching facilities, database querying, and an elaborated and precise layout. The aim of the two following sections is to present in concrete terms what kind of investigations may be done within such a framework, in the context of ancient Greek studies and current digital resources.

[8] For instance, the TEI provides means to encode gaps, additions, deletions, notes, emphasis, etc.

2. Corpus Example: Plato's Work, from Perseus to TXM[9]

2.1 Choice of the texts

As a first experiment of TXM on ancient Greek literature, we have chosen to study a corpus of Plato's texts, because one of us has a good knowledge of these texts, especially of *Gorgias* (Marchand and Ponchon 2016). Actually, the textometric approach presupposes some acquaintance with the corpus to be investigated, because this approach does not compute an analysis on behalf of the researcher: one cannot give the corpus as input, run a (black box) program, and get what would be the complete textometric analysis for the corpus. On the contrary, it is the responsibility of the researcher to lead the investigation and to raise relevant questions so that the produced results can really be meaningful and fruitful. Since *Gorgias* is a rather small, well studied text, we already have some concrete ideas about its features before starting the textometric analysis, so we can check the results of our textometric analysis against our prior knowledge. Furthermore, the software allows one to perform queries that are impossible for a human reader, or at least that would consume a large amount of time. At any rate, we could take this software as an empirical tool that allows us to test some hypotheses, or to give us some clues to continue investigating such an intriguing dialogue as *Gorgias*, and Plato's work more widely.

We downloaded the XML-TEI edition of the Burnet edition of Plato's texts, published by the Perseus Digital Library[10] and provided under a Creative Commons Share Alike 3.0 license. For this experiment, among the 36 available texts, we excluded the works that are generally considered as *spuria* or *dubia* and, for stylistic reason, the letters.[11] Hence the corpus is composed of 29 texts: *Euthyphro*, *Apology*, *Crito*, *Phaedo*, *Cratylus*, *Theaetetus*, *Sophist*, *Statesman*, *Parmenides*, *Philebus*, *Symposium*, *Phaedrus*, *Alcibiades 1*, *Alcibiades 2*, *Charmides*, *Laches*, *Lysis*, *Euthydemus*, *Protagoras*, *Gorgias*, *Meno*, *Hippias Major*, *Hippias Minor*, *Ion*, *Menexenus*, *Republic*, *Timaeus*, *Critias*, and *Laws*.

---

[9] The import process followed here and the resources associated with it are available on txm-users wiki: https://groupes.renater.fr/wiki/txm-users/public/perseus. The page which is dedicated to this article (https://groupes.renater.fr/wiki/txm- users/public/perseus_201707_plato) provides also the binary (.txm) version of Plato's corpus, which can be directly loaded into TXM: this binary version allows one to skip the technical stages of the import process.

[10] Perseus Home: http://www.perseus.tufts.edu/hopper; Text repository: https://github.com/PerseusDL/canonical-greekLit.

[11] Most of the *Letters* are considered as spurious. Moreover the *Letters* belong to a different text type and would disturb contrastive analysis within the corpus (we could study them separately, in another corpus).

2.2 From Perseus TEI encoding to TXM

When we prepared this corpus in June and July 2017, TEI encoding of Plato's texts in Perseus was heterogeneous. We had to deal with texts in several states: last updates made in 2017, 2015, 2014, and 1992. 2017 texts were clearly a new encoding generation. Two texts (*Ion* and *Republic*) used a particular encoding scheme, for instance regarding the use of the element and the marking of the sections' beginning. We decided not to modify sources (which are continuously evolving and improving thanks to the Perseus community), but to make automatized and limited changes during the import process so as to get a usable corpus, even if the TXM user had to contend with some inherited heterogeneity (this is especially the case for speaker information). These automatic changes were processed with XSL and CSS stylesheets,[12] used as parameters in the XML-XTZ[13] import.

The main stylesheet, applied in the second stage of the import process, manages some XML-TEI features of Perseus texts (about nested or <text>) in order to make them compliant with TXM processing (especially for the CQP search engine component embedded in TXM). It automatically gets information from <teiHeader> and associates it with each text instance in TXM: <title>, <author>, and <editor> from <titleStmt> (first mention for each element), and also the content of @when attribute for first (or most recent) <change> element in <revisionDesc>.[14]

Another XSL stylesheet builds the default references provided for word matches in KWIC concordance view: we chose to show the usual name for the text followed by the Estienne page number (Figure 3). Additionally, we ensured that CTS-URN information, providing unique and standard identifiers to cite digital textual data, is also available at the word level, and can alternatively be chosen for references if needed (Figure 4). The main stylesheet was also used to add page break <pb> elements (with the page number) so as to have the same pages in TXM as in the Perseus edition.

---

[12] We are very grateful to Alexei Lavrentiev (CNRS, IHRIM, Lyon), who helped us to adapt the stylesheets to our corpus and needs.

[13] XTZ stands for XML-TEI Zero, that is, this import is intended to process XML-TEI files, taking into account the semantics of some common elements (<text>, <p>, <ab>, <lg>, <head>, <lb>, <pb>, <w>, <graphic>, <ref>, <note>, <hi>, <emph>, <list>, <item>, <table>, <row>, <cell>) if they occur (no mandatory element except <text>). By default, other elements are kept for the analysis and have no rendering in corpus edition view, but this can be fine-tuned by means of XSL and CSS parameter stylesheets. More information is available online in TXM documentation: developer specifications (https://groupes.renater.fr/wiki/txm-info/public/import_xtz), user manual (https://txm.gitpages.huma-num.fr/txm-manual/importer-un-corpus-dans-txm.html#module-xml-tei-zerocsv-dit-aussi-xtzcsv-ou-xtz-import-de-xml-tei-generique-.xml).

[14] A normalized form for short title (title1) and date (update10) is also provided by a metadata.csv parameter file.

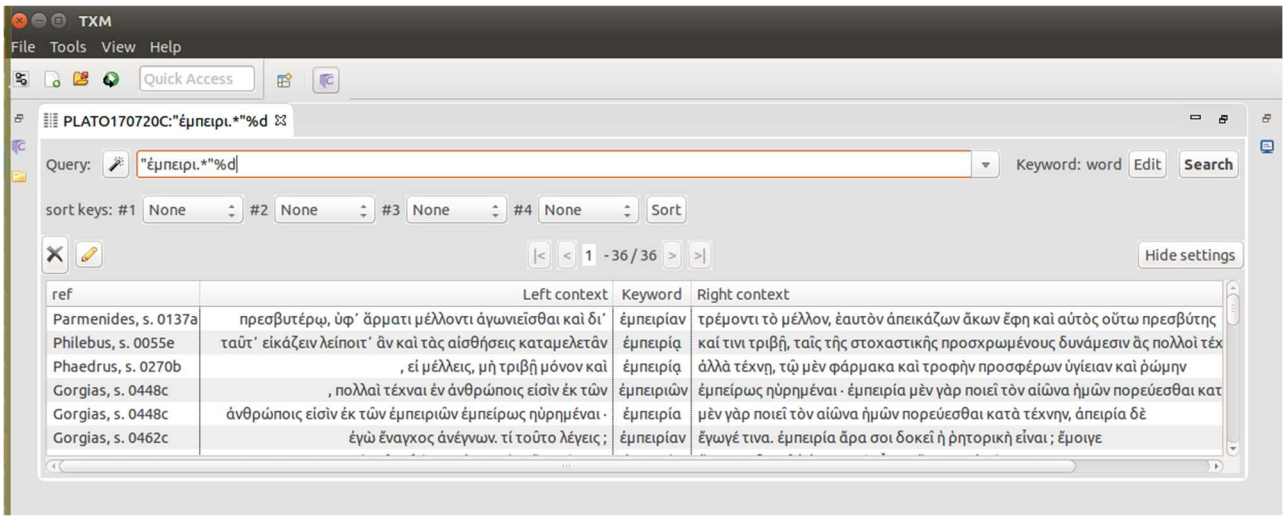


Figure 3. Concordance view with default references in the first column (short title followed by Estienne page number).[15]

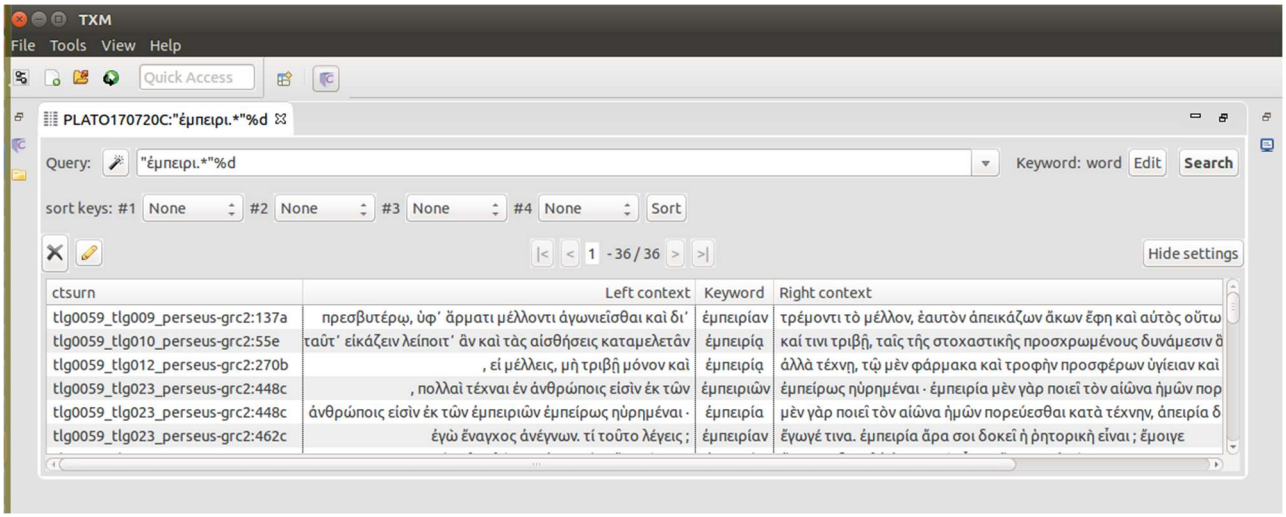


Figure 4. Concordance view with references set to CTS-URN (first column).

We wanted to keep and show clearly speech turn information and speaker information, without indexing the protagonist's name as a word to be counted and searched as such. To this aim we declared <speaker> and <label> elements as “Out-of-text-to-edit” element in import parameters. As text encoding in Perseus was heterogeneous, speaker information is also heterogeneous in TXM. Speech turns are defined by either <sp> or <said> element (depending on the text), and speaker is indicated with @who attribute only in <said> elements (but not all of them) for analysis purposes, and is displayed using <label> or <speaker> element content when available (Figure 5).

[15] The query is: "ἐμπειρι.*"%d. It matches any word beginning with ἐμπειρι, and the %d operator allows diacritic variations, so that this query matches ἐμπειριῶν but also ἐμπειρίᾳ with an accent on the last ι.

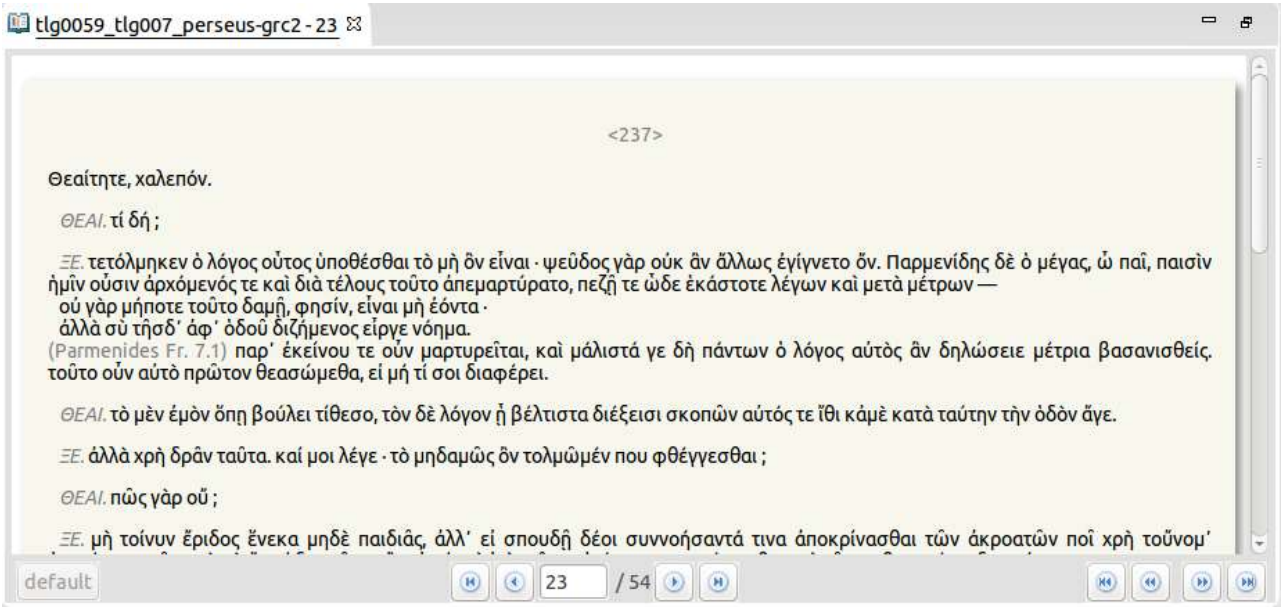


Figure 5. TXM edition view of *Sophist*, Estienne page 237a-e.

We should just stop a minute on that: it means that with this edition of Perseus corpus in TXM, one can search the word Γοργίας and differentiate between the instances of this word in the title (obviously one instance), in speaker names (97 instances in 1 dialogue),[16] and in the speech of any protagonist (85 instances in 7 dialogues);[17] one could also study which words are typical of Socrates' or Gorgias' speech. This functionality opens up a lot of possibilities, namely, to study the particular style of each protagonist of the platonic dialogue.

Bibliographic references encoded with <bibl> element in Perseus texts are declared and processed as a kind of note element for TXM XML-XTZ import: as such, they get a relevant display (see Figure 5) and are not mixed with the Greek content.

Every annotation made in Perseus texts, for instance, named entities (<name>, <persName>, <placeName>) or quotation tags (<q>, <quote>), are automatically available in TXM for search and analysis. An associated rendering could be added through the CSS stylesheet parameter if desired.

In addition, in order to test new possibilities offered by morphosyntactically tagged texts, we prepared a second corpus using the unique text of Plato available in Perseus TreeBank AGDT2:[18]

---

[16] Example of query to find every beginning of Gorgias' speech turns:
<said>[_.said_who="#Γοργίας" & _.said_rend!="merge"]

[17] Example of query to find mentions of Gorgias' name in the text: "Γοργί.*". The 7 dialogues with "Gorgias" occurrences are *Apology*, *Gorgias*, *Hippias Major*, *Meno*, *Phaedrus*, *Philebus*, and *Symposium*. Since we had no morphosyntactically tagged version of our texts, we use wild-cards to get a rough lemmatization. This query matches Γοργία, Γοργίας, Γοργίαν, Γοργίου, Γοργίᾳ.

[18] Ancient Greek Dependency Treebank AGDT2: https://perseusdl.github.io/treebank_data.

Euthyphro. We imported this text in TXM via the XML/w loader with the Perseus TreeBank stylesheet[19] as parameter. This corpus will be used here for one example in section 3.4.

3. A Typology of Textometric Analyses

For our example, we focus on the text of *Gorgias*, in comparison with Plato's other dialogues. So we define two structures inside the corpus: first, a *sub-corpus* containing only the *Gorgias* text, so that any textometric analysis (i.e. a KWIC concordance, a frequency list, etc.) can be processed on this text only. The second structure that we define is a *partition* of the corpus into texts, that is, we divide the corpus into parts (texts) so that we can compute contrastive analyses in order to compare texts to one another.

We have chosen to organize the typology according to the user's needs, that is, which kind of queries one would like to ask the corpus. These query types don't exactly match textometric functions, since one computation may be of use in several ways, and one type of query may get an answer through the complementary results of several functions.

3.1 Checking for occurrences and evaluating frequencies

To look at word frequencies in *Gorgias*, we can either compute raw frequencies, or compare the *Gorgias* subcorpus to the whole corpus of Plato using a statistic test, computing the specificities' score of each word (more precisely, each form of the word, since our corpus is not lemmatized) (see technical appendix 5.1).

Thus, for *Gorgias*' vocabulary, we can sort the results either by absolute frequency, or by specificities' score. In Figure 6, the left panel shows the most frequent words used in *Gorgias*, which here is useless since the most frequent words of *Gorgias* are the most frequent words in the corpus of Greek literature: they are "stop-words" like καί, δέ, and so on…[20] If we apply a stop-word filter,[21] we get the most frequent content words of the text, giving account for the main vocabulary used in *Gorgias* (see middle panel).

---

[19] txm-filter-perseustreebank-xmlw.xsl, available in TXM repository: https://sourceforge.net/projects/txm/files/library/xsl/.

[20] To see a list of the ancient Greek stop-words, cf. https://wiki.digitalclassicist.org/Stopwords_for_Greek_and_Latin.

[21] There is no direct command in TXM to do this. If the corpus is tagged with part-of-speech, then this information makes it easy to select and filter out grammatical words (see example in Figure 18). Here, we downloaded Berra's stop-word list (https://github.com/aurelberra/stopwords); we used the CreateCQPList TXM macro (https://groupes.renater.fr/wiki/txm-users/public/macros) to read the stop-word file and build a variable containing all the stop-words (about 6,000 items); then we called this variable ("grc_stopwords") in the query to

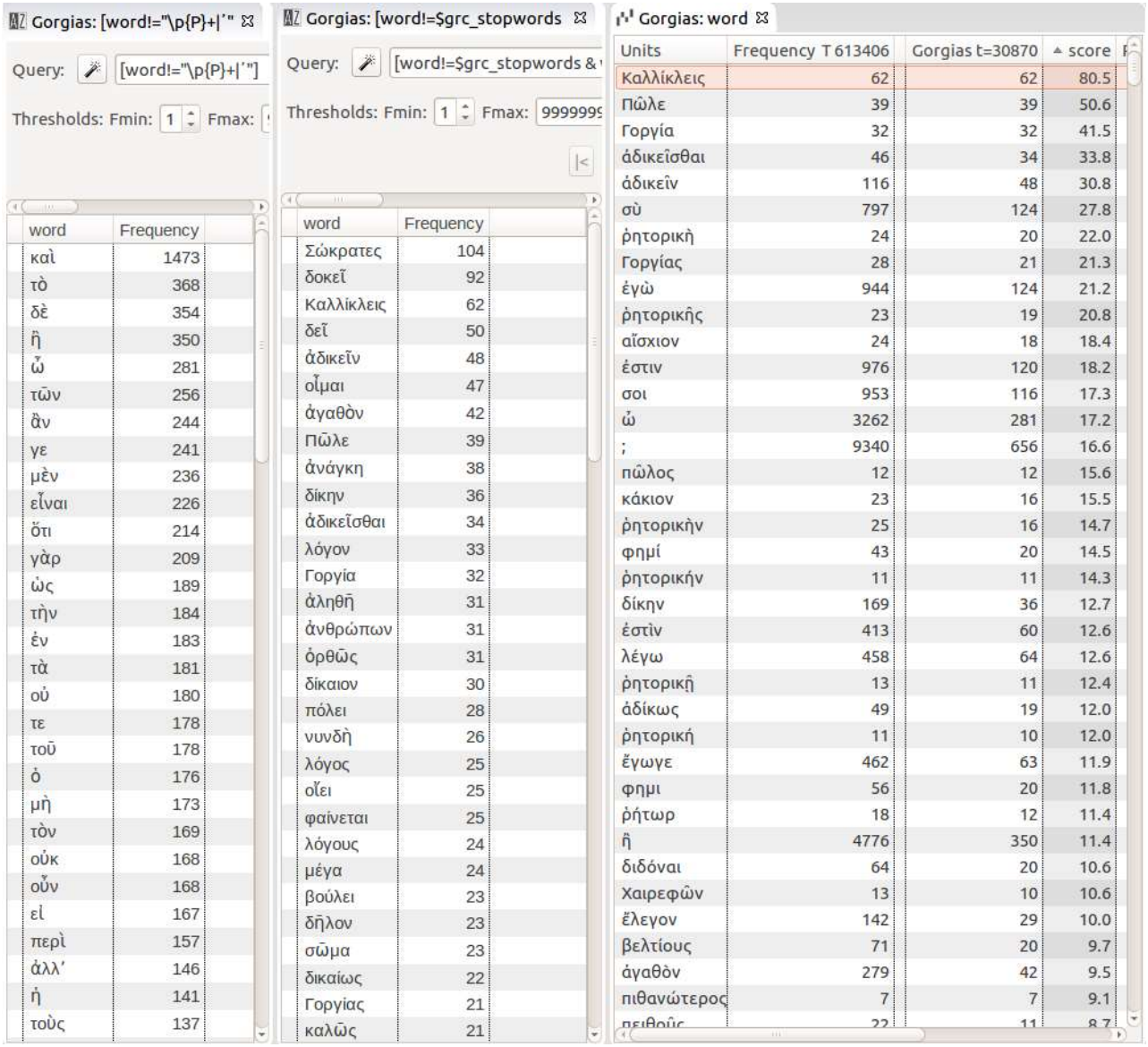


Figure 6. For *Gorgias*, the list of most frequent word forms[22] (left panel), the same list without stop-words (middle panel), and the statistically most specific word forms (right panel).

The right panel may be more interesting, as it shows the words statistically over-used in *Gorgias* by comparison with the 28 other texts from Plato. The fact that the first three specific forms are Καλλίκλεις, Πῶλε, and Γοργία, that is, names of the protagonists, was expected, since those three protagonists do not appear elsewhere in Plato's work, and then are naturally specific to Plato's *Gorgias*.

---

exclude words matching any item of the stop- list. As TXM default tokenization preserves elided word forms (like ἀλλ', δ', etc.), for this experiment we just added to Berra's list the elided stop-words we found among the high-frequency words of our corpus. (We might also have changed the tokenization rules.)

[22] TXM indexes punctuations too, so here the query filters punctuation signs out.

More significant is the fact that they are vocative forms, which are highly frequent in the dialog form: *Gorgias* is a dialog that is a real exchange (like *Laches*, *Euthydemus*, *Euthyphro*, *Hippias Major*, for instance), and not the very particular form of platonic dialog where the protagonists do not really interact intensively with each other, which are closer to a series of monologues (like the *Laws*, the *Parmenides*, *Republic*, and the *Timaeus*). This "dialogic" aspect of *Gorgias* became evident at the first sight of the specificities of the dialog: not only the over-use of vocative Σώκρατες, but also such dialog markers as σύ, ἐγώ, ὦ… Those first results show by linguistic means the refutative nature of the *Gorgias*, which is an attempt to refute sophistic positions and to expose the immoral implications in the hedonist position defended by Callicles. For Socrates, this requires repeatedly questioning the interlocutor (using the vocative), summarizing his argument (by using σὺ + verbs of saying φής, ὡμολόγεις, ἔλεγες, λέγεις…: "you says, you agree, you said..."), and opposing it with his own position (ἐγὼ λέγω, ἔλεγον…) (see also Figure 16 in section 3.4). Admittedly, *Gorgias* is not the only dialog where we can find those markers; they are common to what the scholarship calls the "Socratic dialogs," which entails an ἔλεγχος, a refutation. But regarding those features, the textometric approach shows that *Gorgias* is representative, if not the dialog *par excellence* (Figure 7). It reveals that this dialog is paradigmatic, even something of an achievement of this method, since the dialog suddenly ends with a monolog and the acknowledgment that it is impossible to refute those who do not agree with the basic requirements of a rational discussion; that is why it ends with a myth. This analysis confirms the hypothesis that *Gorgias* is a transitional dialog, where Plato shows the limits of the Socratic method and the necessity to endorse another philosophical method.[23]

[23] On this hypothesis, see (Marchand and Ponchon 2016:17–18).

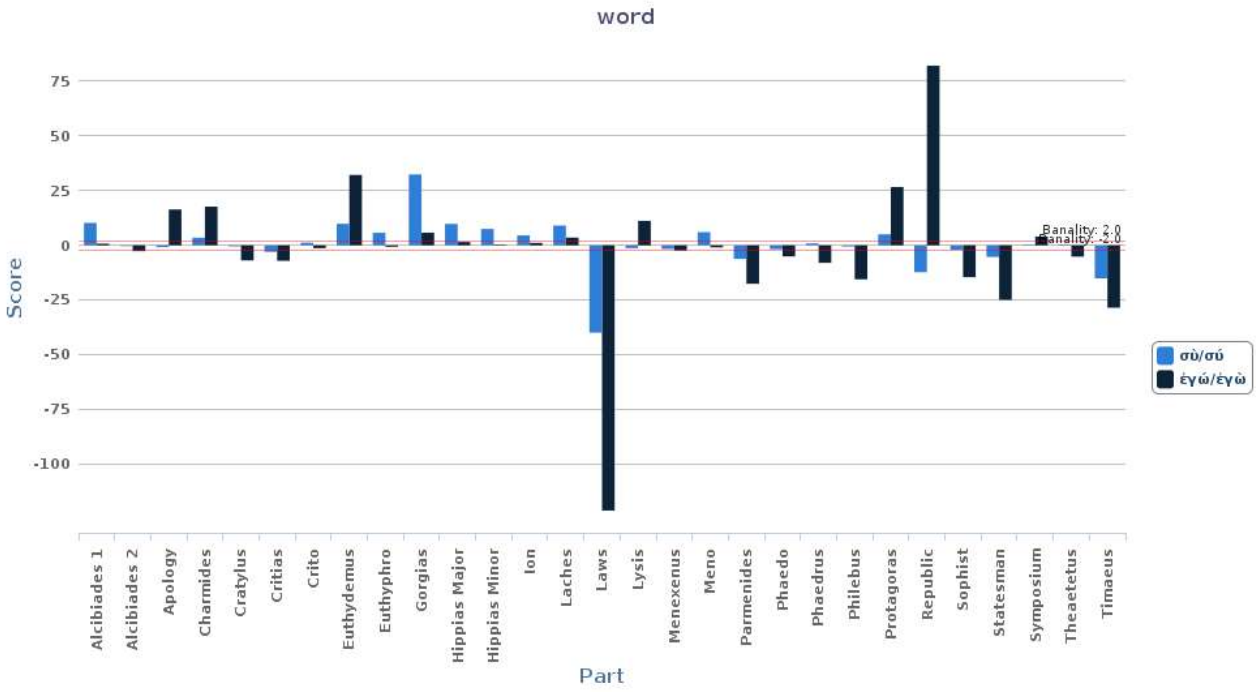


Figure 7. Specificities of ἐγώ and σύ in the texts of the Plato corpus.

3.2 Visualizing the evolution of words or linguistic features

The TXM progression functionality is another way to analyze word distribution when data can be considered as a continuous evolution: this can be the case for a corpus when its texts are ordered according to time, or when focusing on one text, considering how it is organized from beginning to end. This second case can be applied to the *Gorgias*.

After the dialogic markers, the list of the specificities above shows also what are roughly the main themes of the dialog. The subtitle (which is probably not from the hand of Plato) of *Gorgias* is “on the rhetoric” (περὶ ῥητορικῆς), and it is well known that the main topic is a critique of the practices of the ῥήτορες, the men who deliver rhetorical speeches and teach rhetoric. Despite this very fact, the word ῥητορική (more precisely, all the forms of the adjective ῥητορικός)[24] appear mostly in the first quarter of the dialog (there are 107 instances in *Gorgias*, and 85 appear from 448d to 466a, that is, in the discussion with Gorgias himself) (Figure 8).

[24] Since the corpus is not lemmatized, we submitted the query "ῥητορικ.*" (Figures 8 to 12 have been computed with this query).

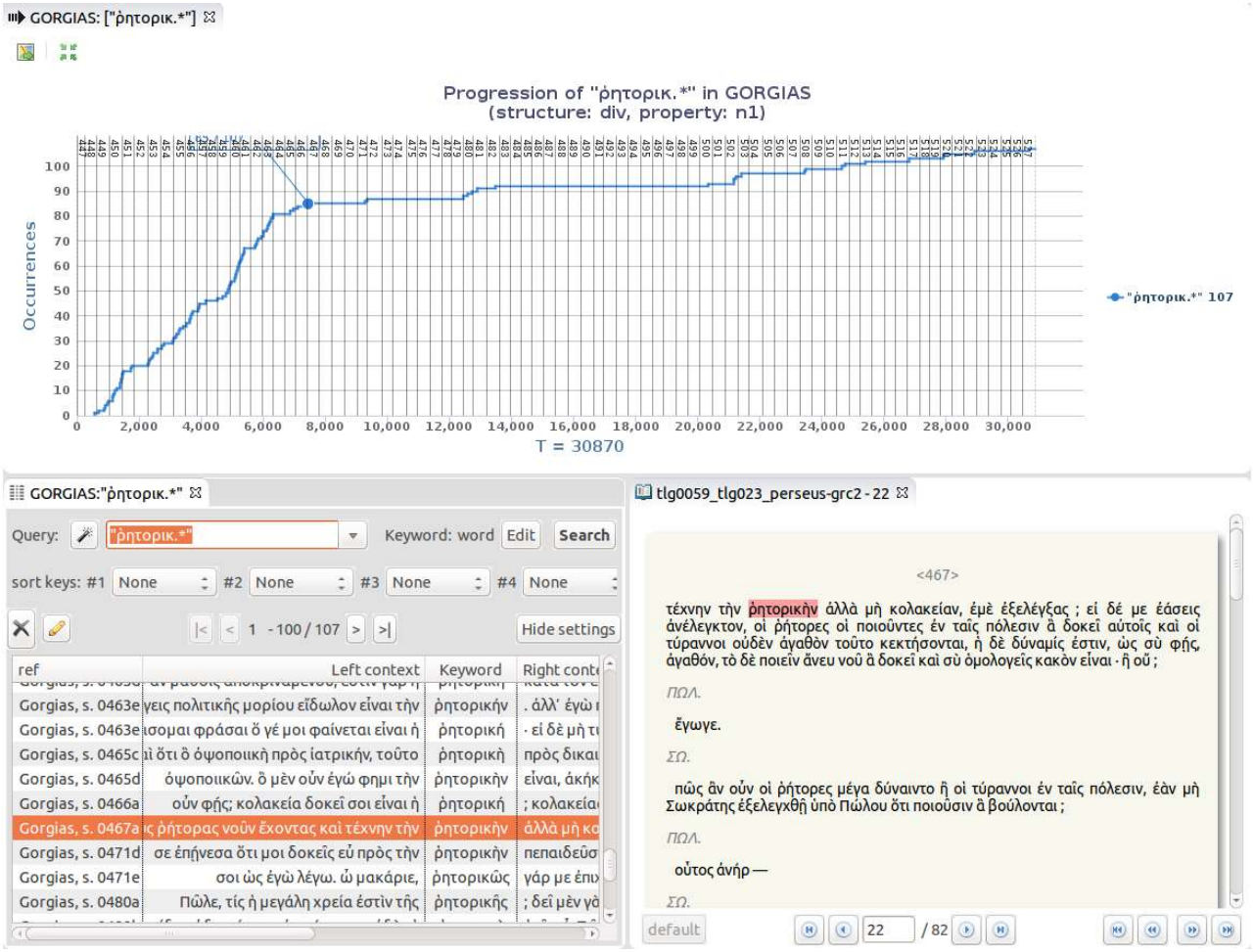


Figure 8. Cumulative evolution graph for ῥητορική in *Gorgias*[25] (upper frame), and hyperlinked concordance and text view on a selected occurrence (lower frames).

This is due to the fact that the Socratic criticism of rhetoric is a moral criticism of the way of life involved in the practice of the rhetoric. Socrates shows that the practice of rhetoric, as the faculty to persuade a crowd in order to take power, involves an immoral principle, namely, that it is better to do wrong than to suffer it. On the contrary, the philosophy of Socrates lies in the belief that it is better to suffer wrong than to do it (see *Gorgias* 469c, the question from Polos: εἰ δ᾽ ἀναγκαῖον εἴη ἀδικεῖν ἢ ἀδικεῖσθαι, ἑλοίμην ἂν μᾶλλον ἀδικεῖσθαι ἢ ἀδικεῖν; "if it were necessary either to do wrong or to suffer it, I should choose to suffer rather than do it," translation Lamb from Perseus). For that reason, the forms ἀδικεῖσθαι and ἀδικεῖν are among the most specific forms of the *Gorgias* (they appear just after the three vocatives already

[25] X-axis represents the *Gorgias* text from its first word to its last word; Y-axis represents the number of occurrences of words beginning by ῥητορικ- since the beginning of the text. Thus, the more the curve rises, the more occurrences there are in the current passage. A flat curve means no occurrences of the word in the passage.

mentioned Καλλίκλεις, Πῶλε, and Γοργία), although they are in general quite common in Plato's work.[26]

A glance at the table of frequencies and graph of specificities of the adjective ῥητορικός (all its cognates) in every text of our corpus (Figure 9) shows, however, that ῥητορικός is typical of the *Gorgias*, for the very reason that it is not very common in Plato's work: hence, even if Plato does not continually use the word in the dialog, it is nonetheless characteristic of the *Gorgias* (107 instances, specificity of +108), and to a lesser extent the *Phaedrus* (21 instances, specificity of +8) (these texts are the only two where the specificity score is greater than +3).[27]

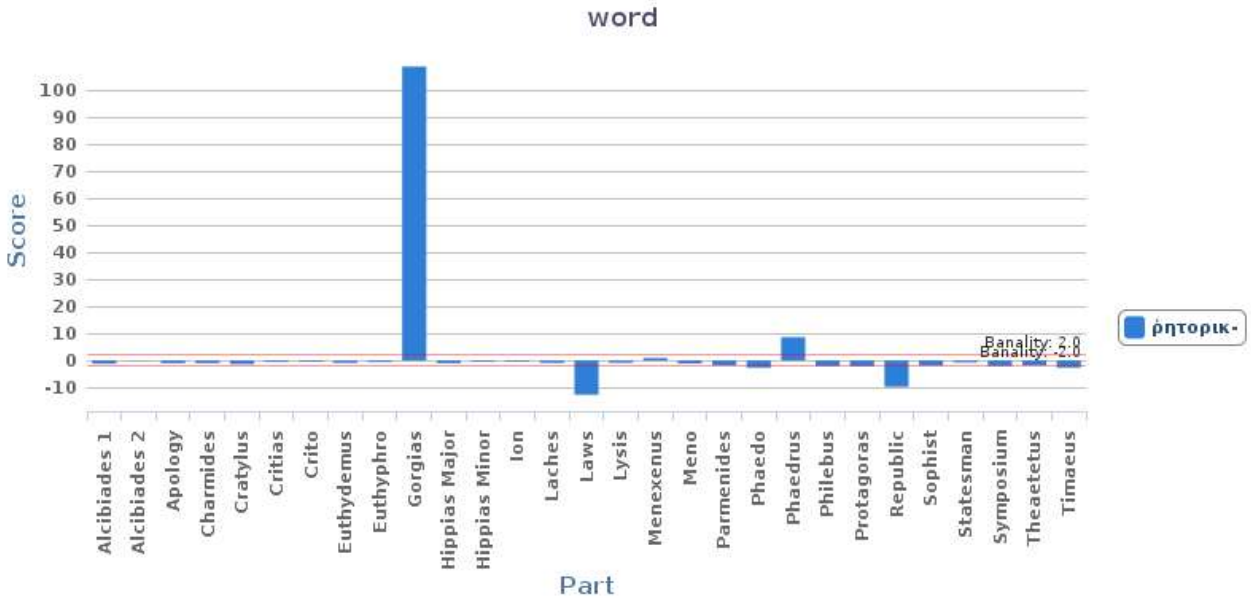


Figure 9. Specificities of ῥητορική in the texts of the Plato corpus.

E. Schiappa has polemically stated that it could be possible that Plato coined the word ῥητορική.[28] This is not the place to take a position on the external evidence discussed by Schiappa, who claims that all the texts that employ the word ῥητορική in the fourth century can be considered contemporary or later than Plato.[29] But the tool here allows us to confirm the fact that in Plato "the various forms of ῥητορική are curiously distributed." (Schiappa 1990:463) and

---

[26] 46 instances in Plato of ἀδικεῖσθαι (34 in *Gorgias*); 116 instances of ἀδικεῖν (48 in *Gorgias*). The query "ἀδικ.*" on whole of Plato's work shows, with no surprise, a massive presence in *Republic* (164 instances), *Gorgias* (163 instances), and *Laws* (103 instances); the presence in *Republic* and *Laws* has to be relativized by the size of the dialogs (respectively 104,572 and 116,925 words, whereas for *Gorgias* there are 30,870 words).

[27] See the other instances, *Statesman* 304e, 304d; *Republic* 548e; *Menexenus* 235e, 236a (bis), *Euthydemus* 307a, *Cratylus* 425a, *Alcibiades 2* 145e, *Thaetetus* 177b.

[28] Schiappa 1990. See also the criticisms of O'Sullivan 1993 and the revised Schiappa 1999.

[29] By the facts, the dates of those texts are difficult to determine, as is the case for Alcidamas or the *Rhetoric to Alexander*.

the graph shows that this curious distribution relies on the specificity of *Gorgias*' vocabulary (Figure 9).[30]

3.3 Refining a word's meaning with systematic contextual use

The KWIC concordance view is a core functionality in the textometric approach, as it shows very precisely and efficiently how a word is used in the text. The context size can be adjusted if needed, but for a view of a larger selection of text, a double-click on a concordance line opens a hypertext link to the corresponding text page with the search word highlighted. The two views are dynamically aligned: selecting a line in the concordance shows the corresponding page in the text edition (Figure 10).

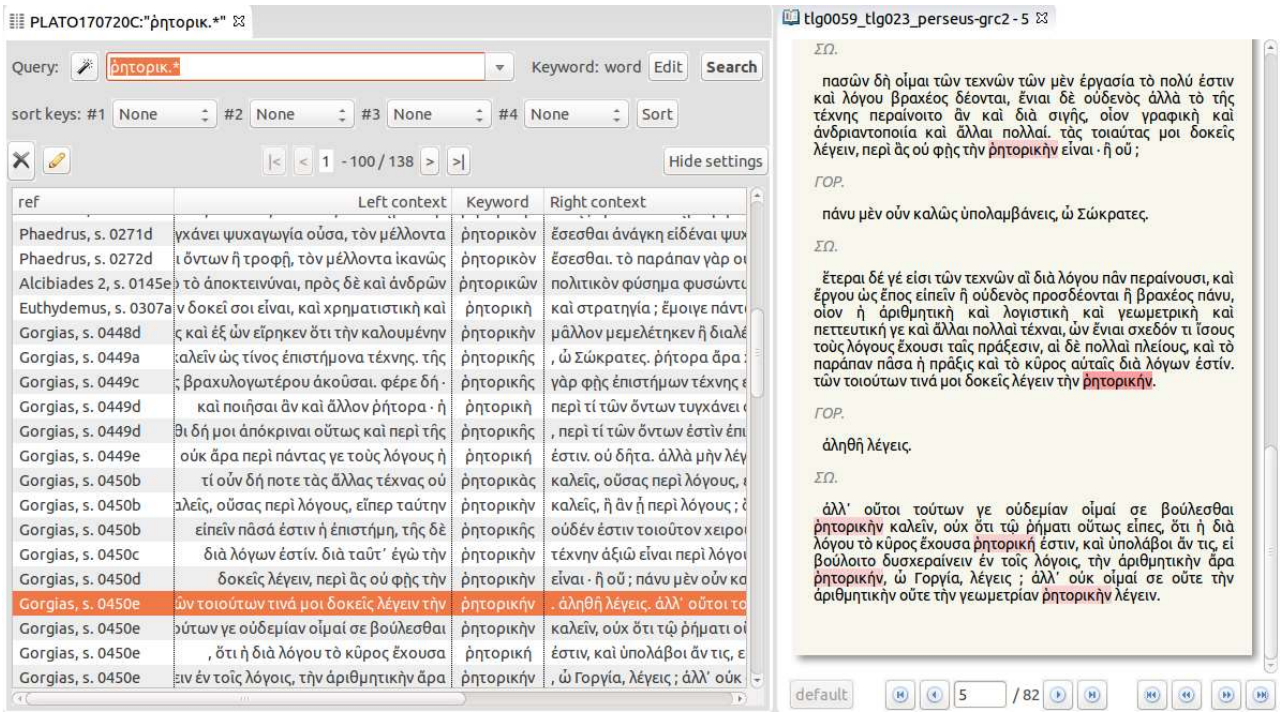


Figure 10. Concordance of ῥητορική in the Plato corpus, and the hyperlinked view of the text page corresponding to one selected concordance line.

Furthermore, the table layout of the KWIC view, coupled with the possibility to sort right and left contexts, reveals the lexical patterns involving the searched word. For instance, sorting on the 'right context' of ῥητορική allows one to find the passages where the concept is defined: all the contexts where ῥητορική is followed by the verb εἶμι (to be) or καλέω (to call) (Figure 11).

[30] By the way, among the 30 first words with the highest score of specificity in *Gorgias*, there are 6 forms of ἡ ῥητορικὴ.

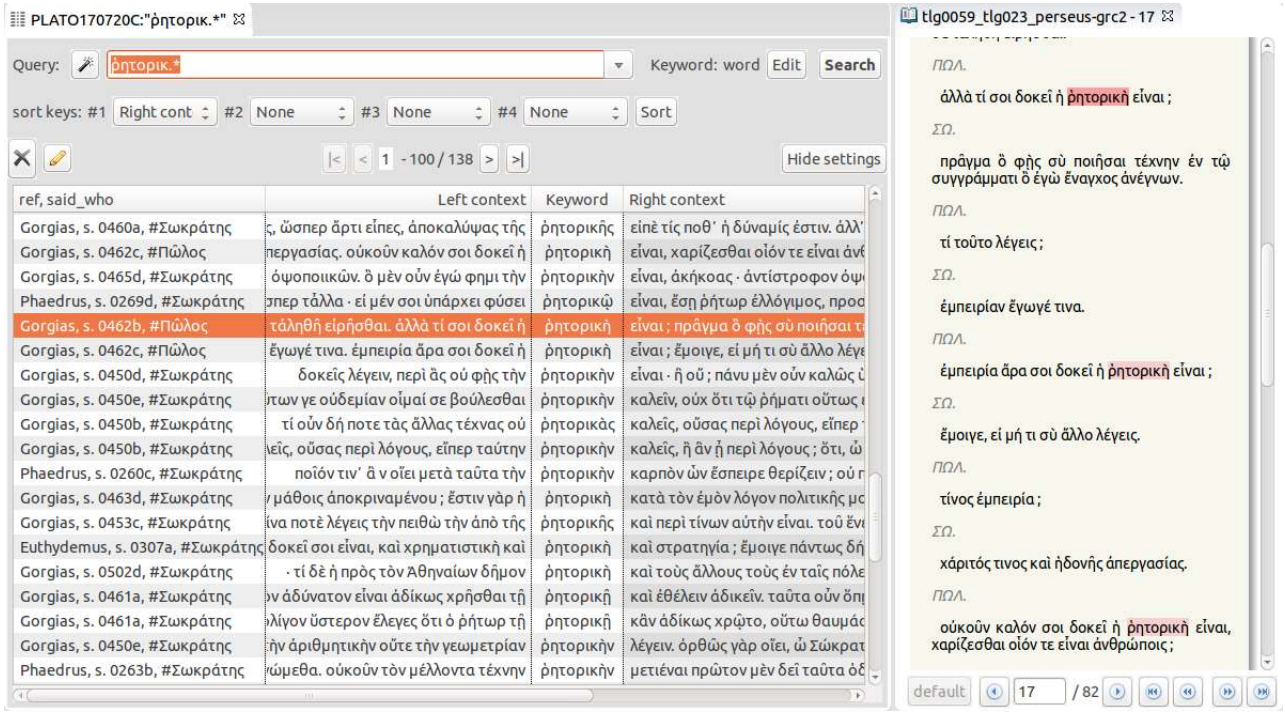


Figure 11. Concordance of ῥητορική in the Plato corpus, sorted on 'right context'.

Lastly, to have an overview of the platonic approach of the word ῥητορική, one can have a look at its cooccurrents[31] in the *Gorgias*: this query shows the words especially used by Plato in proximity with ῥητορική and its cognates (Figure 12). The cooccurrence association score is computed with the same statistic measure as the specificity explained above (the cooccurrence score is precisely the specificity of the cooccurrent word in the part formed by the whole set of contexts of ῥητορική in comparison with its global frequency in the corpus). The common feature of the platonic definition of rhetoric appears within a range from -10 to +10 words:[32] rhetoric is the name of the competence of the ῥήτωρ; Socrates defines it as "producer of persuasion" (e.g. 453a: πειθοῦς δημιουργός, 453d: πειθὼ ποιεῖν); its definition is given through an analogy as part (μόριον) of the art of flattery (κολακεία; e.g. 463b); it is used in the context of the court (δικαστηρίοις)… It is worth noting that those cooccurrents are also the main cooccurents of the word ῥητορική in all Plato's works.

[31] What is called "cooccurrence" in textometry is often called "collocation" in corpus linguistics.

[32] This context size is a common choice, but the software allows for the exploration of different settings parameters if needed.

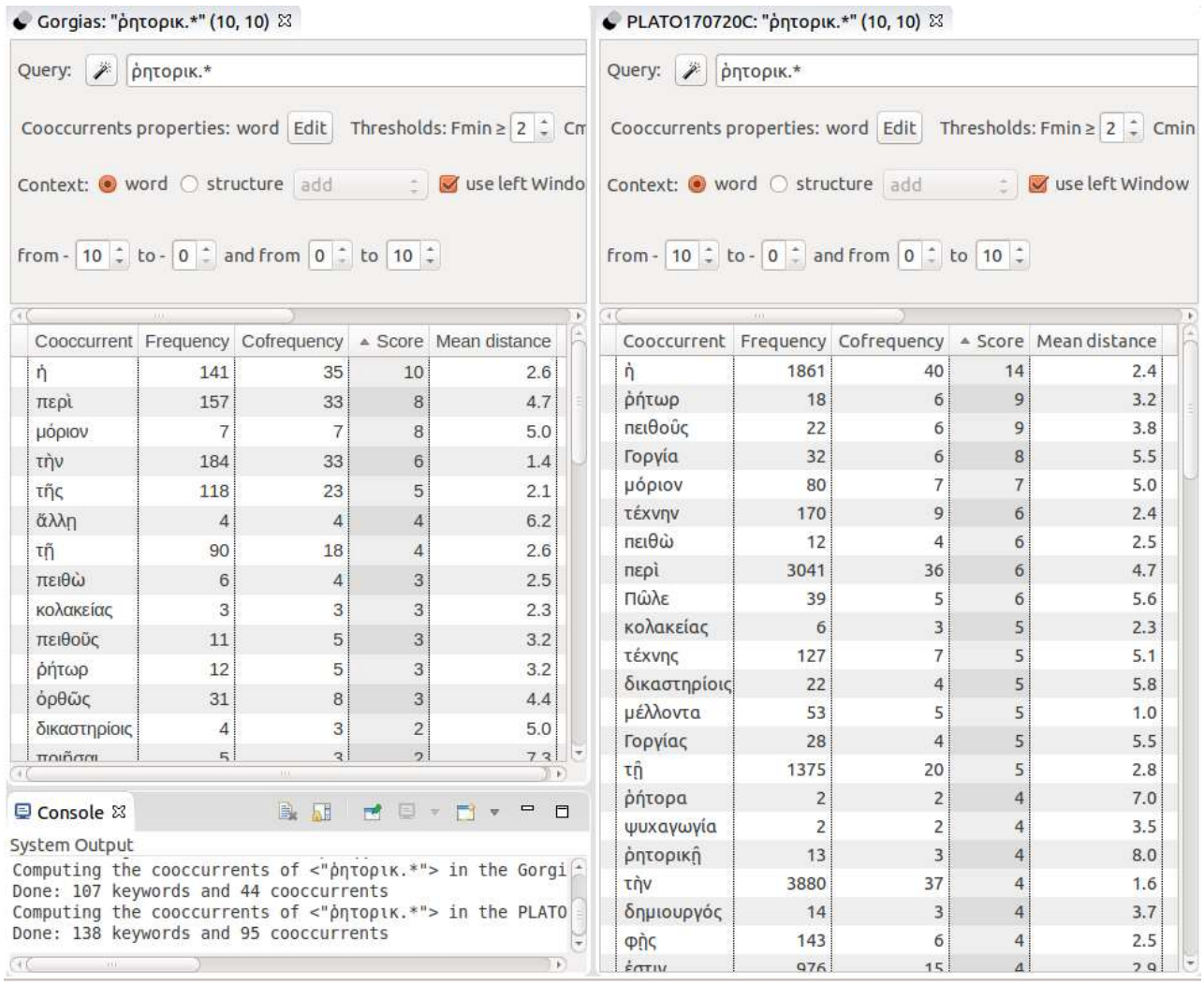


Figure 12. Cooccurrences of ῥητορική in *Gorgias* (left panel) and in the whole Plato corpus (right panel).

In the same fashion, the query of the cooccurrents of ἀδικε.* (which matches both ἀδικεῖσθαι and ἀδικεῖν) (Figure 13) instantly points to the passages where Plato makes the comparison between suffering and doing wrong. It is no surprise that the cooccurrents are ἀδικεῖσθαι (because it is most of the time compared to ἀδικεῖν), ἀδικεῖν (for the same reason), τὸ (because the comparison is between the fact of doing or suffering wrong), αἴσχιον, and κακίον (because in the dialog with Polos, the question is mainly whether it is fouler or more evil to commit or to suffer wrongdoing).

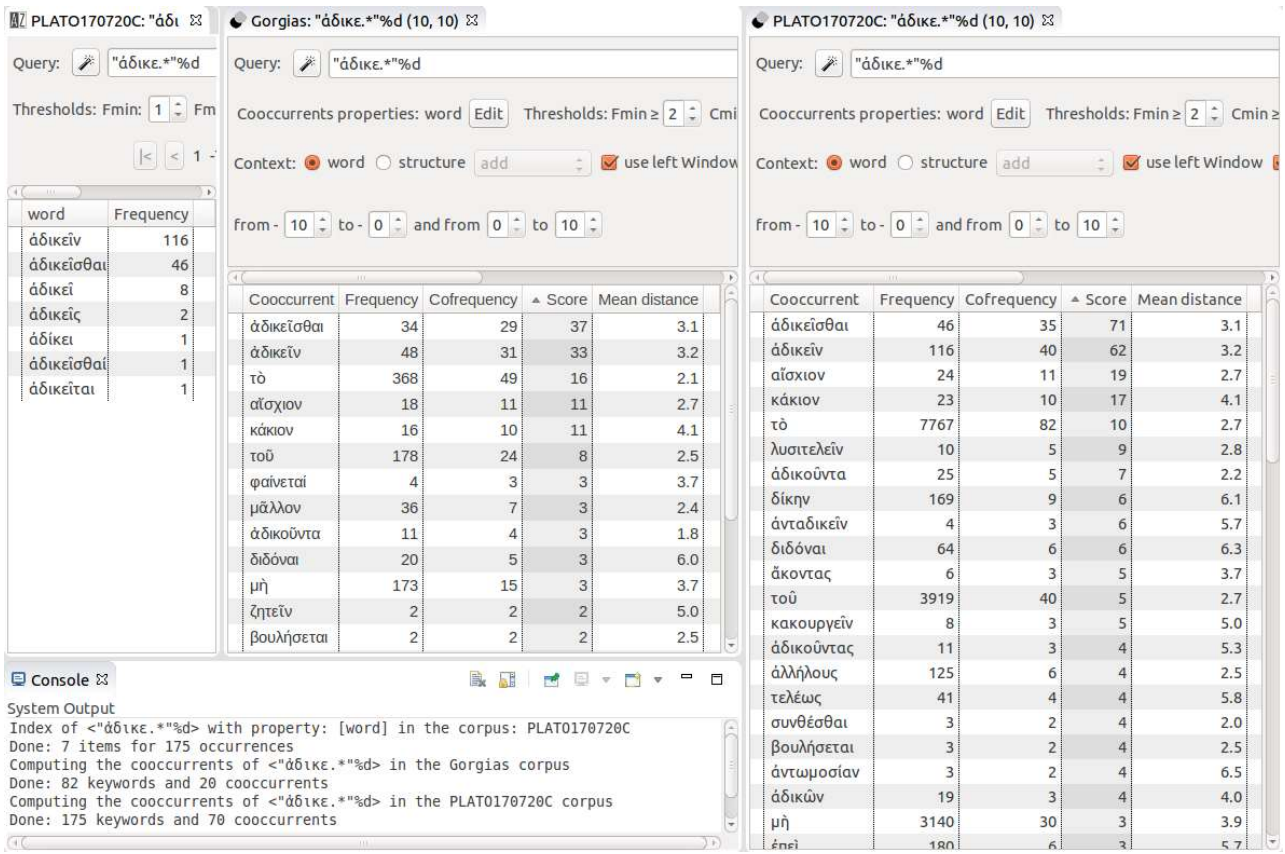


Figure 13. Cooccurrences of ἀδικεῖσθαι and ἀδικεῖν (cf. complete word forms list on the left) in *Gorgias* (central panel) and in the whole Plato corpus (right panel).

An interesting case arises when cooccurrences in the *Gorgias* happen to be quite different from the cooccurrences in the whole corpus, that is, instances in which a word gets a distinct meaning in the *Gorgias*. This is the case for νόμος (Figure 14): in the *Gorgias* ὁ νόμος is mainly used in the discussion with Callicles in the context of the opposition between what is κατὰ νόμον ("according to convention") and κατὰ φύσιν ("according to nature"). In the other dialogs (and mainly in the *Laws*) the cooccurences of νόμος emphasize the description of what will be the law, or conform to the law (ἔστω κατὰ νόμον), or the political action of the law (κείσθω νόμος).

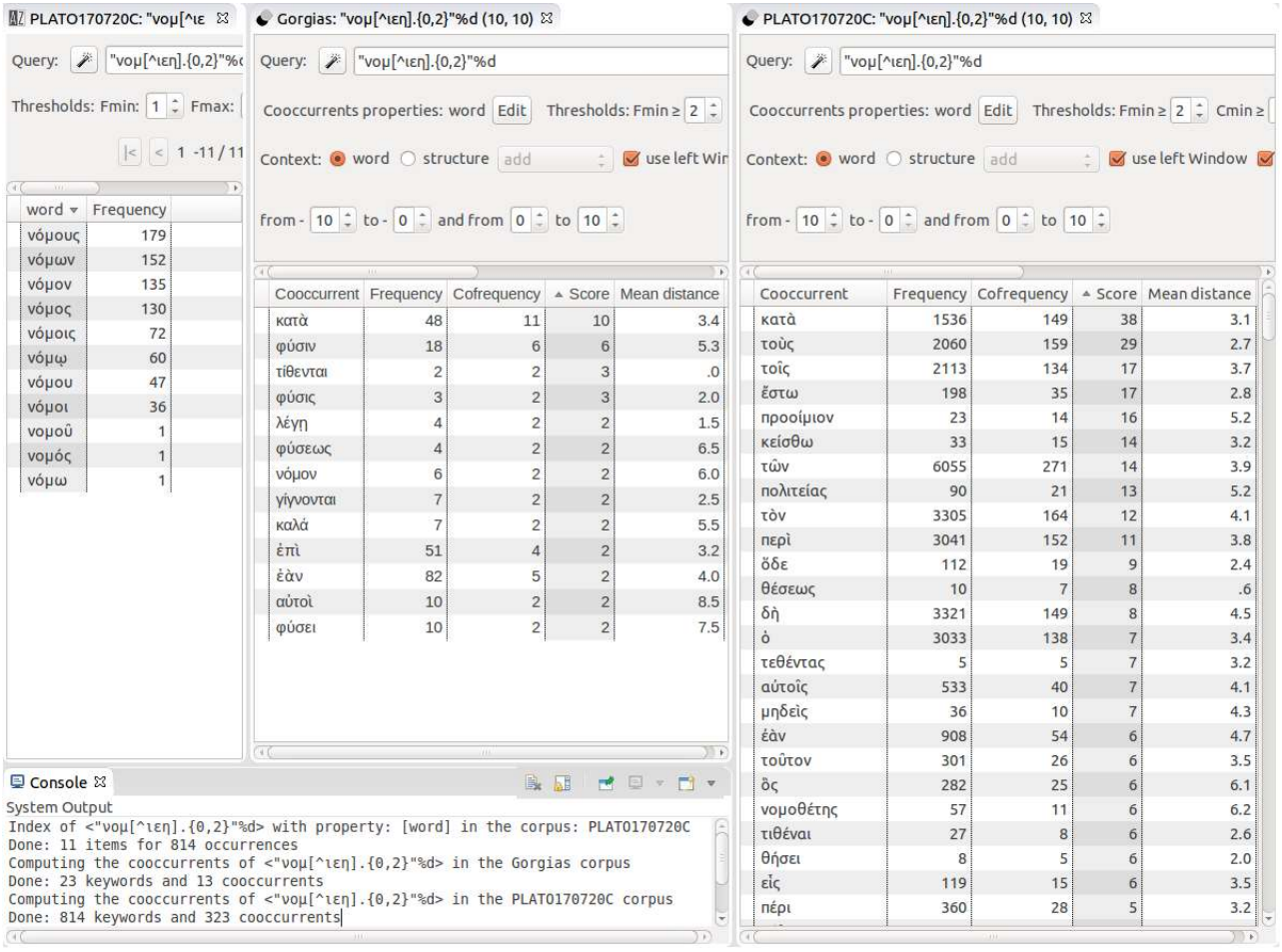
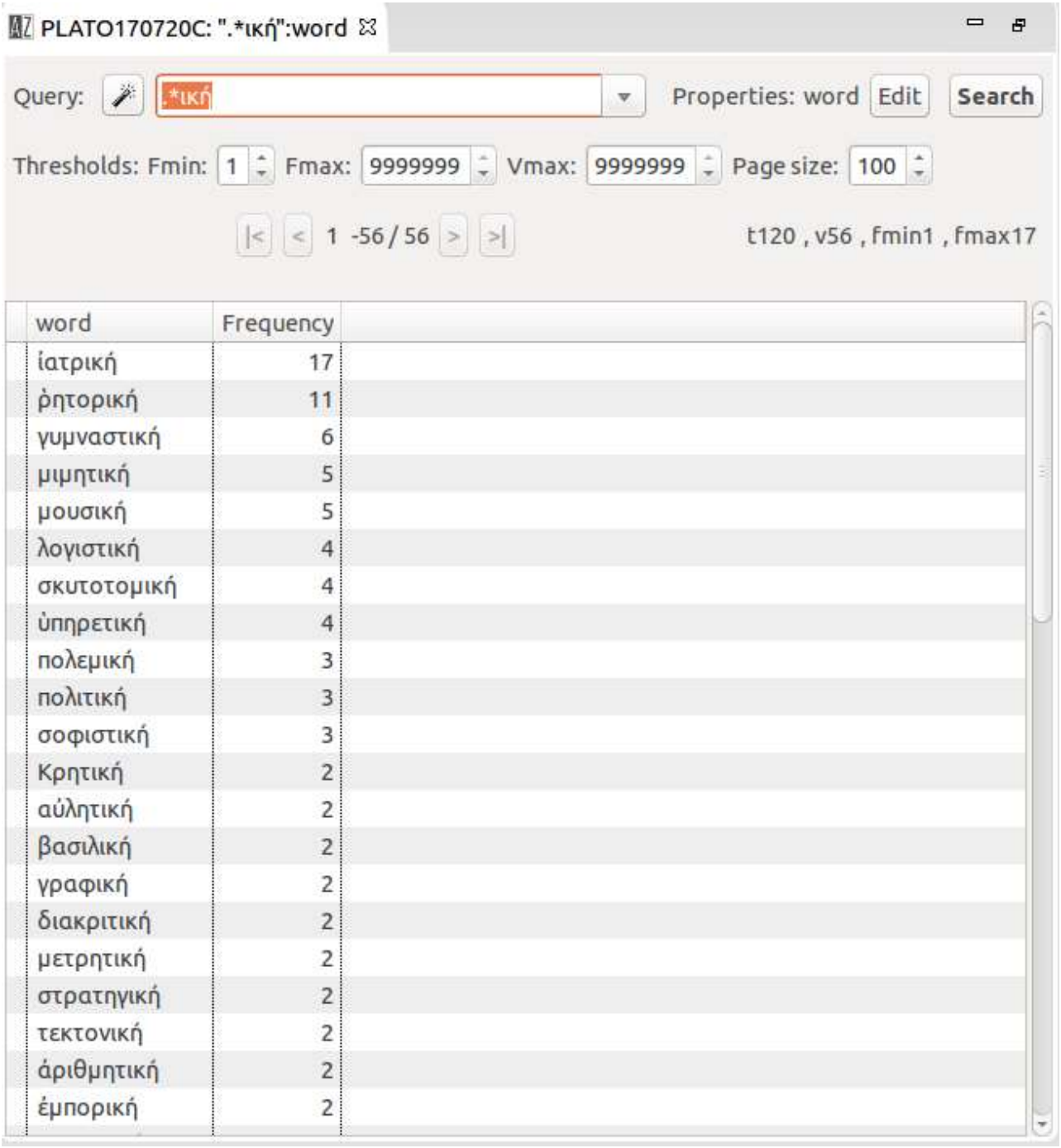


Figure 14. Cooccurrences of νόμος (cf. complete word forms list on the left) in *Gorgias* (central panel) and in the whole Plato corpus (right panel).

3.4 Computing corpus-based paradigmatic series

Another way of using the search engine is to build a list of words or patterns that share some constitutive or contextual feature. For instance, one could list all the lexical units including the same given morphological unit, or list all units occurring in some precise position (words ending verses, adjectives qualifying a given noun).

Schiappa emphasized “Plato was a prolific coiner of words ending with -ική denoting ‘art of’” (Schiappa 1990:464). We can have a look at our data to get the set of these kinds of words (Figure 15).



word	Frequency
ἰατρική	17
ῥητορική	11
γυμναστική	6
μιμητική	5
μουσική	5
λογιστική	4
σκυτοτομική	4
ὑπηρετική	4
πολεμική	3
πολιτική	3
σοφιστική	3
Κρητική	2
αὐλητική	2
βασιλική	2
γραφική	2
διακριτική	2
μετρητική	2
στρατηγική	2
τεκτονική	2
ἀριθμητική	2
ἐμπορική	2

Figure 15. Index for the query .*ική in the 29 texts of Plato, results for a minimum frequency of 2.

In section 3.1 above, we noted that in the *Gorgias*, Socrates frequently uses the pronoun σύ with verbs of saying, and ἐγώ with verbs expressing his own position. We can get a rough but systematic summary of verbs more frequently used with ἐγώ and σύ in the Gorgias by computing their cooccurrences in a narrow context window (5 words right and left) (Figure 16).

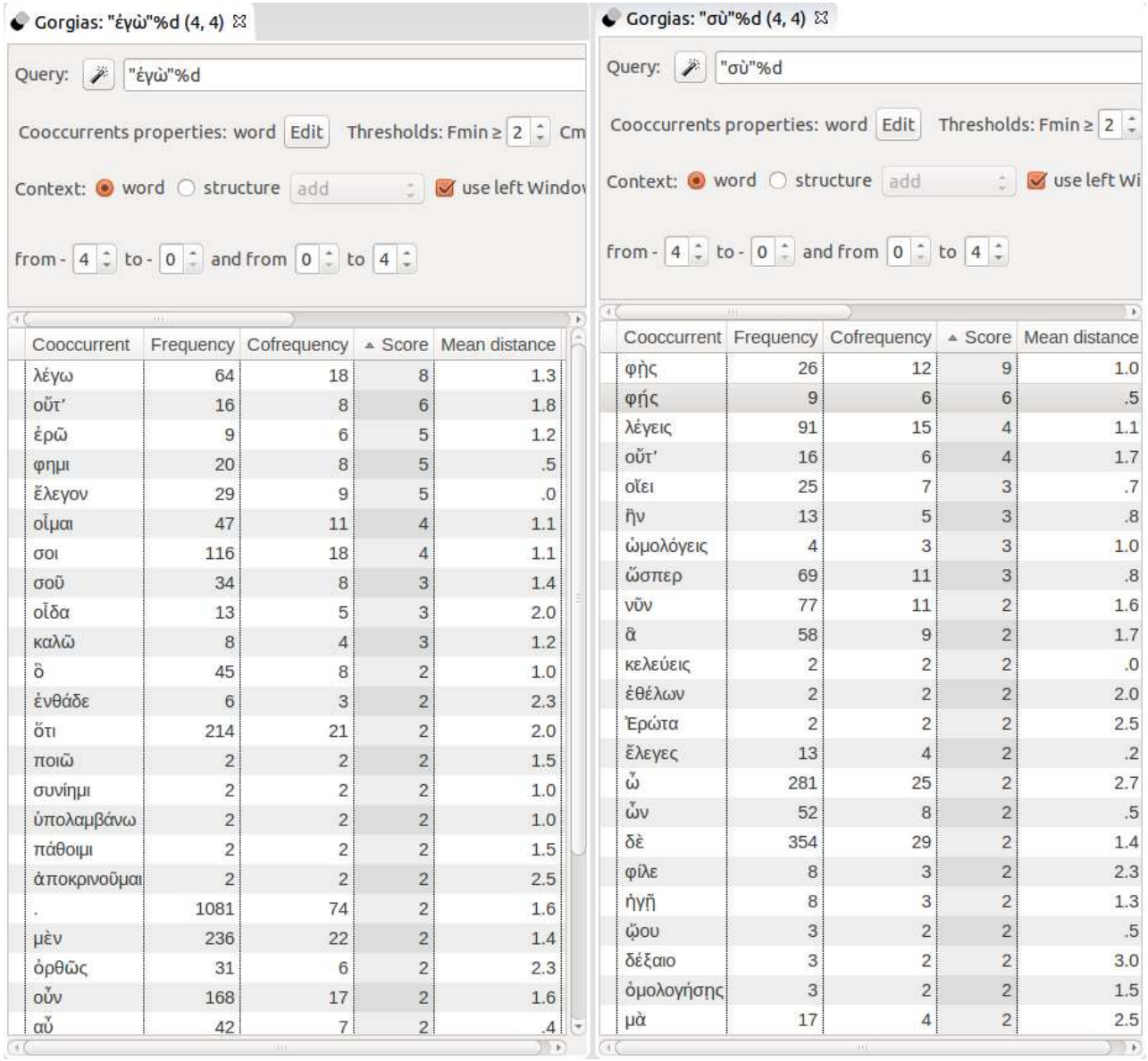


Figure 16. Cooccurrences of ἐγώ and σύ in Plato’s *Gorgias*.

Another kind of paradigmatic set can be computed on the basis of textual structures. For instance, we can have a look at every one-word answer in the corpus, and thus explore the different ways in Plato's dialogs to express agreement or disagreement (Figure 17).

PLATO170720C: (<said>[][]</said> | <sp>[][]</sp>):word

Query: (<said>[][]</said> | <sp>[][]</sp>) Properties: word Edit Search

Thresholds: Fmin: 1 Fmax: 9999999 Vmax: 9999999 Page size: 100

|< < 1 -100 / 172 > >| t1377 , v172 , fmin1 , fmax430

word	Frequency
ναί .	430
ἔγωγε .	87
πῶς ;	72
φαίνεται .	68
οὕτως .	50
ὀρθῶς .	47
ἀνάγκη .	43
ἔοικεν .	41
οὐδαμῶς .	32
ἀληθῆ .	30
φημί .	27
ἀληθέστατα .	27
ἔμοιγε .	26
ἴσως .	21
πῇ ;	20
ὀρθότατα .	17
κινδυνεύει .	16
καλῶς .	15
οὐδέν .	14
δῆλον .	12
ἔστιν .	12

Figure 17. Index of one-word answers.

But the Plato TXM corpus we are using here does not reveal all the possibilities given by this kind of query. A more precise investigation could be done on a tagged corpus. Let us have a quick look to two other examples taken outside of *Gorgias*.

Our first example deals with morphosyntactic annotation. The *Euthyphro* corpus was built from the TreeBank AGDT2 and has morphosyntactic tags, so we can list the verbs occurring in the same sentences as πατήρ (Figure 18).

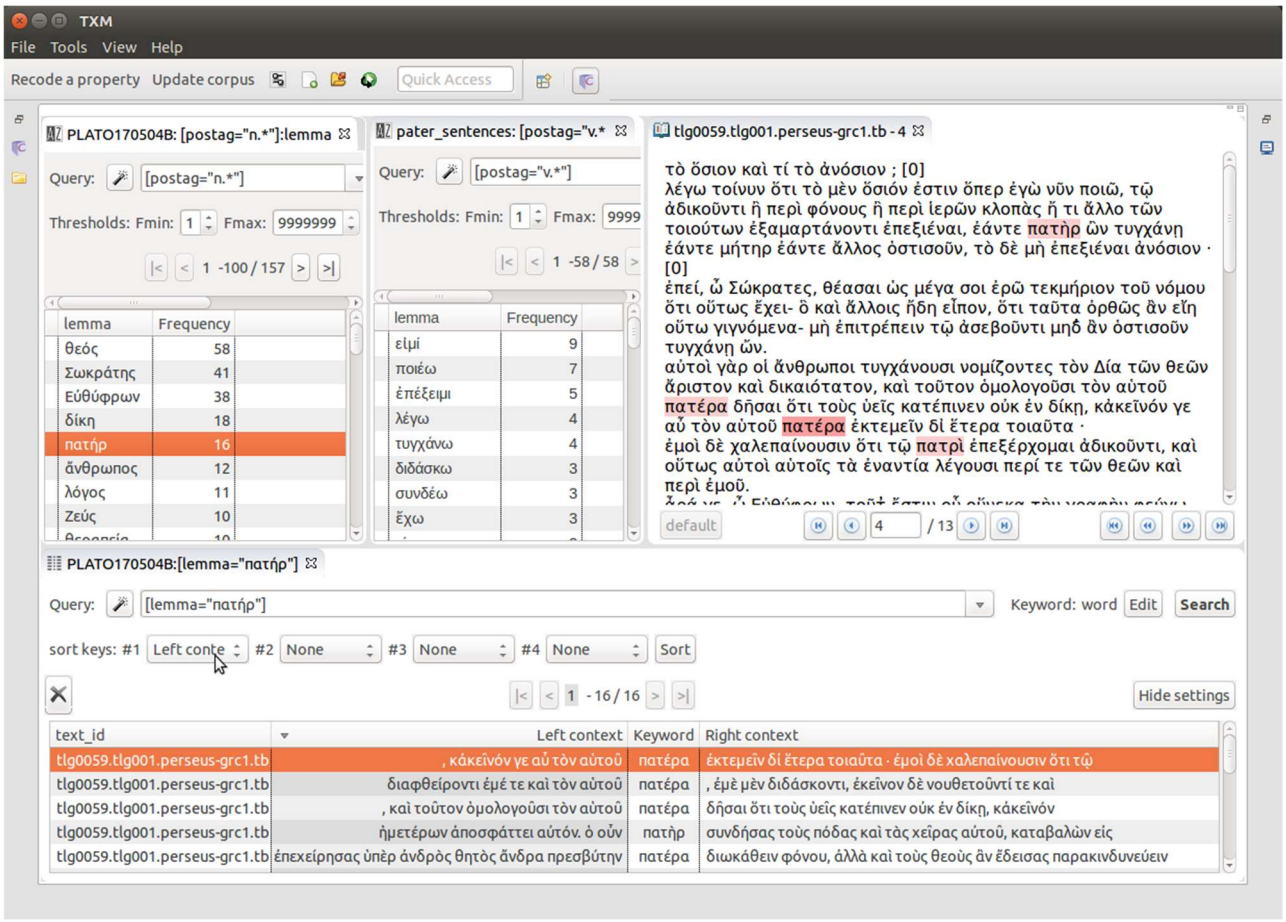


Figure 18. Verbs used with πατήρ in *Euthyphro*. (i) (upper left) search for most frequent nouns in the text (lemmas); (ii) (lower) contexts of πατήρ in a concordance view; (iii) (upper right) widening context in a page view; (iv) (upper central) list of verbs (lemmas) in sentences containing πατήρ, sorted by descending frequency.

Obviously, all encoded information can be used to define what will be returned. Our second example uses semantic information. In our Plato corpus, one text (the *Phaedo*) has been tagged for named entities; we can use this annotation to see precisely which places or people are mentioned in *Phaedo* (Figure 19).

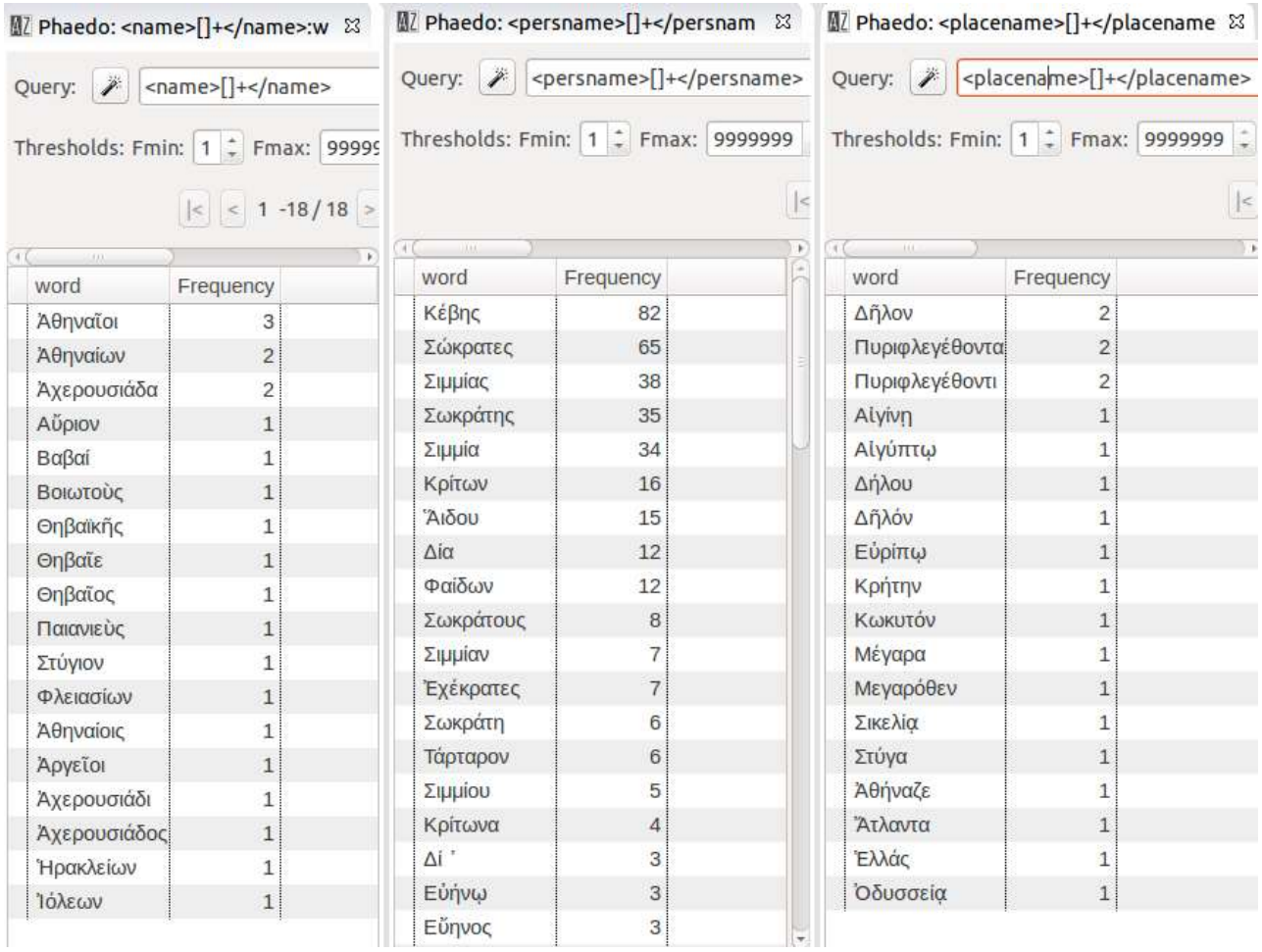


Figure 19. Index of named entities encoded in the Perseus edition of *Phaedo*.

3.5 Local contrastive analysis of a corpus: Identifying what is typical in a part

Studying the vocabulary of *Gorgias*, we have already shown that the specificity measure can be used according to two points of view: for a word (ex. ῥητορική, ἐγή, σύ) we can draw its usage profile among texts; and for a text, we can point out the words that are statistically overused in this text in comparison to the rest of the corpus. This way, it is possible to study and compare every division of the corpus. For example, if we have a hypothesis that a corpus is divided into four main parts (be it text types, time spans, streams, etc.), we have a tool to automatically extract words that could help grounding and exploring this hypothesis.

As a complement to statistical processing, another way to look at the particularities of the *Gorgias*' vocabulary is by simply listing the words that appear only in Gorgias[33] (Figure 20).

[33] In TXM, the table used is the one produced by computing the specificities on the *Gorgias* subcorpus

Gorgias: word

Units	Frequency T 613406	Gorgias t=30870	score	PLATO170720C \ Gorgias
Καλλίκλεις	62	62	80.5	0
Πῶλε	39	39	50.6	0
Γοργία	32	32	41.5	0
πῶλος	12	12	15.6	0
ῥητορικήν	11	11	14.3	0
πιθανώτερος	7	7	9.1	0
Καλλικλῆς	6	6	7.8	0
Πῶλον	6	6	7.8	0
ὀψοποιικὴ	5	5	6.5	0
κόρρης	4	4	5.2	0
ἐλέγχου	4	4	5.2	0
ἐπιψηφίζειν	4	4	5.2	0
ἐροῦ	4	4	5.2	0
δημηγορία	3	3	3.9	0
διδούς	3	3	3.9	0
ηὑρίσκομεν	3	3	3.9	0
κάεσθαι	3	3	3.9	0
προσφέρει	3	3	3.9	0
ἀπαλλαττόμενος	3	3	3.9	0
ἀποκτενεῖ	3	3	3.9	0
ἀφαιρήσεται	3	3	3.9	0
Ἀρχέλαος	3	3	3.9	0
ἐάσεις	3	3	3.9	0
ἔλεγές	3	3	3.9	0
ἔλεγχε	3	3	3.9	0
ὀψοποιία	3	3	3.9	0
ὁμολογήσῃς	3	3	3.9	0

Figure 20. Word forms used only in Plato's *Gorgias*, with a minimum frequency of 3 (null frequency in all other texts of the Plato corpus).

For instance, putting aside proper names, the word-form ἐπιψηφίζειν (to put a question to the vote) has 4 instances in Plato, all of them in *Gorgias*—to which we should add one instance of ἐπιψηφίζων (476a). Even if the term is rather technical, *prima facie* this result might be surprising, given the fact that Plato is known as a critic of Athenian democracy and of the fact that the citizens make political decisions without any expertise, and the verb ἐπιψηφίζειν precisely describes the democratic process of putting a question to the vote in a democratic assembly (see e.g. *Thucydides II* 24). But, if we go back to the text, we can observe that in *Gorgias*

those instances of the lemma ἐπιψηφίζω do not intend to give criticism of the democratic process; rather, they are a reference to the fact that Socrates does know how to put a question to a vote because "there is also one whose vote I know how to take, whilst to the multitude I have not a word to say" (474a). Hence, the choice of ἐπιψηφίζειν in *Gorgias* seems coherent with the project to show the opposition between the rhetorical discourse addressed to the crowd in order to obtain the vote in the Assembly and the philosophical dialog with a sole person in order to decide if the thing in question is true or false. Thus, even if *Gorgias* is full of the insinuation that democracy ruined Athens, those mentions entail, not a real critique of democracy, but an opposition between philosophical and democratic discourse.

Another word form occurring only in *Gorgias* is ὀψοποιική (scil. τέχνη, the art of cookery), which is not a common word in ancient Greek. If we consider all the forms of this word,[34] Plato uses it once in the *Symposium* (187e), when the physicist Eryximachus pronounces that his craft "set high importance on a right use of the appetite for the dainties of the table, that we may cull the pleasure without the disease." But in *Gorgias* the word occurs 7 times, mainly in the well-known passages where he makes a comparison between rhetoric and the art of cookery, precisely to show that neither are arts, but kinds of flattery (Figure 21).

[34] Query: ὀψοποιικ.*

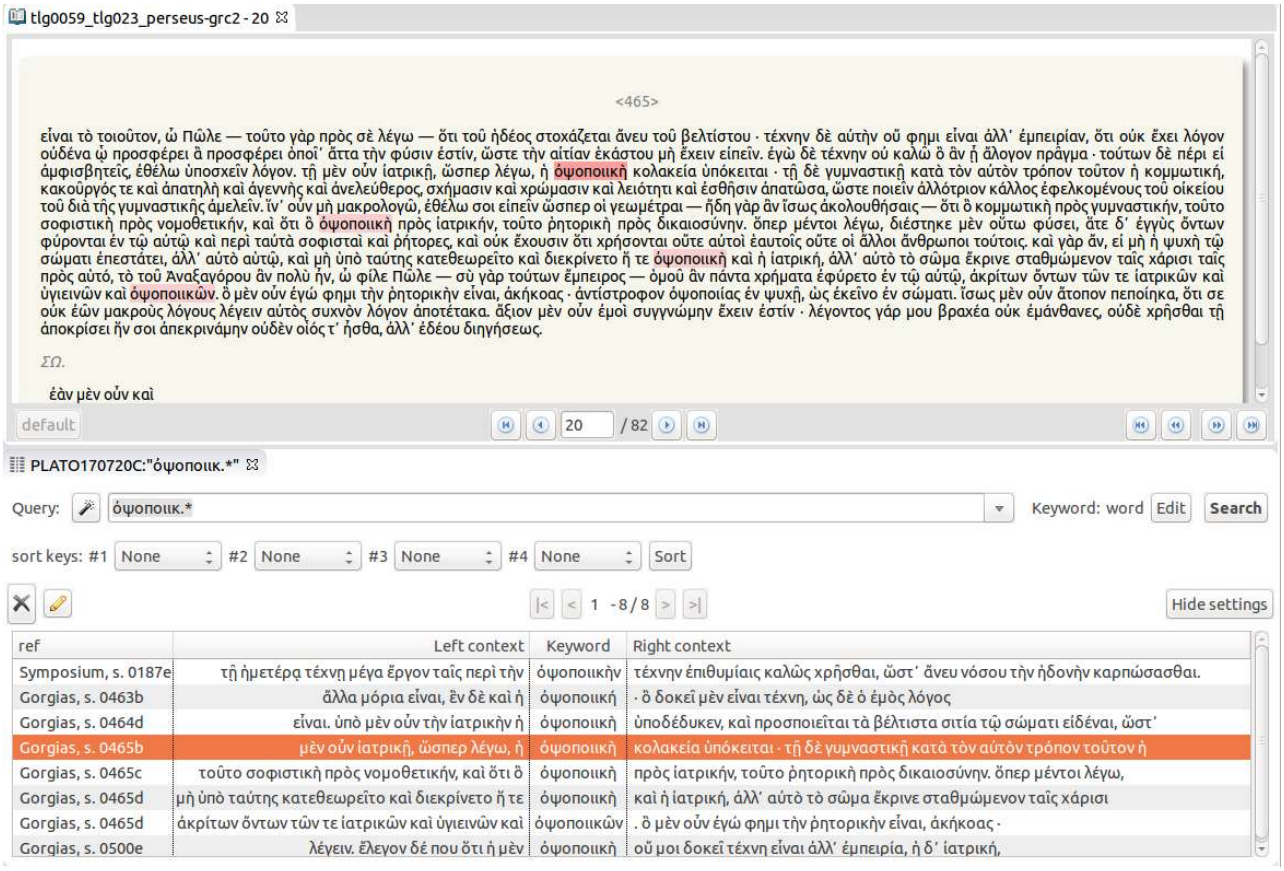


Figure 21. Concordance of ὀψοποιική in the Plato corpus.

This specific use of ὀψοποιική has to be compared with the more common ὀψοποιία and ὀψοποιός, which appear in *Euthydemus* (290b), *Republic* (373c), and *Theaetetus* (178d), but mainly, once again, in *Gorgias* (6 instances, with 3 forms of ὀψοποιός). The frequency of ὀψοποιική in *Gorgias* has to be linked with the strategy already mentioned for the word "rhetoric" and more generally for words ending with -ική (section 3.4). The TLG gives Plato (*Gorgias* and *Symposium*) and Xenophon's *Oeconomicus* as the first instances of the term in Greek. According to the scholarship, the date of composition for Plato's *Gorgias* is 385–380 (Marchand and Ponchon 2016:19), and for the *Oeconomicus* 362; there is also relative agreement that the *Symposium* is later than *Gorgias*. These claims are obviously mere hypotheses because of the lack of textual and external information about the date of composition for those texts. However, it seems that the specificity in *Gorgias* of ὀψοποιική, which appears when Socrates is giving a definition of rhetoric, could be an argument in favor of Schiappa's hypothesis on the platonic origin of rhetoric. It would give also a humorous tone to the definition, if it is true that Socrates compares the emphatic new word to describe the "art" of Gorgias (so-called "rhetoric") with an emphatic new word to describe the "art" of a cook (463b)!

3.6 Overall contrastive analysis of a corpus: Identifying the main dimensions structuring a corpus

Last but not least, for an overall view of the corpus, correspondence analysis is a multidimensional statistical tool, which computes the main dimensions of contrast structuring a corpus (see technical appendix 5.2).

We have generated such an analysis on the Plato corpus. We focused on the 200 most frequent words, which are mainly grammatical words. Each text is represented by the frequencies of its use of these words, which makes a kind of grammatical or stylistic profile (few lexical words). As explained in section 5.2, the x-axis and y-axis represent complex quantitative "mixtures" of words, and they are used to select the best 2D-perspective in the geometric representation of the data. The map (Figure 22) illustrates the relative positions of texts among one another, the main structural associations and oppositions they draw.

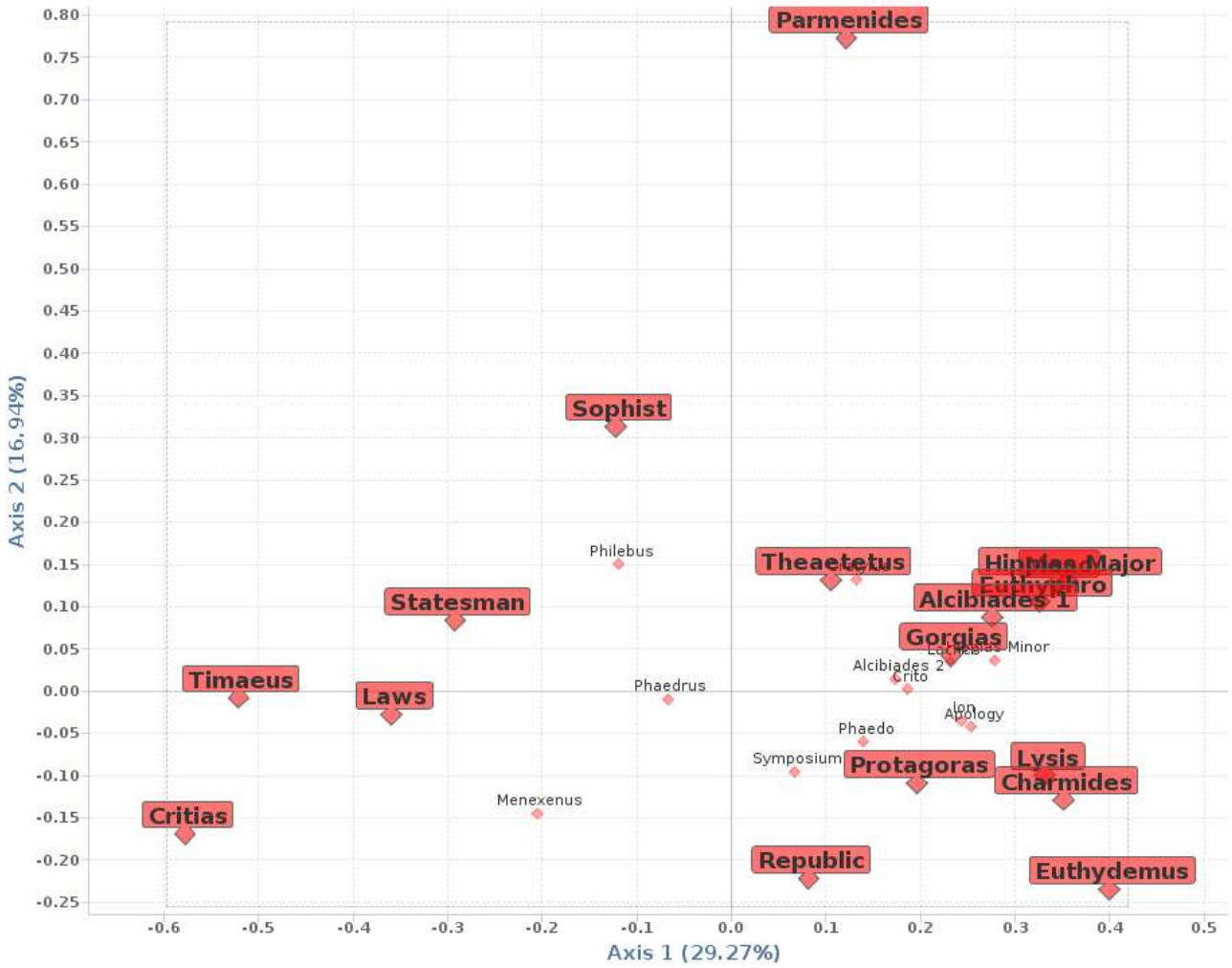


Figure 22. Correspondence analysis map of the 29 texts of Plato characterized by the 200 most frequent words of the corpus.[35]

The main opposition we observe in the corpus is a contrast between, on the right-hand side of the map, dialog and interaction markers (ἔφη, ὦ, ἐγώ, Σώκρατες, ὅτι, μοι, σοι, εἰ, ἀλλὰ, ἔγωγε, οὐ, ὥσπερ, γάρ, πάνυ) and, on the left-hand side, grammatical words more associated with nouns, and then descriptive, written style (δὲ, τῶν, κατὰ, δὴ, τῆς, τὴν, εἰς, τὰς, μὲν, ὅσα, τοῖς). This written style is typical of *Laws*, *Timaeus*, *Critias*, and *Statesman*. Texts are more heterogeneous as concerns dialog and interaction; the main texts concerned are *Euthydemus*, *Gorgias*, *Meno*, *Hippias Major*, *Charmides*, *Lysis*, *Alcibiades 1*, *Protagoras*, and *Euthyphro*. Other texts are less concerned with this opposition; they may mix markers or use few of them.

Other correspondence analyses could produce different insights into the corpus: for instance, we could focus on a lexical characterization instead of a grammatical one, and try to

[35] Red labels emphasize texts that are best represented on this map (their quality indicator, computed as the sum of squared cosine for axes 1 and 2, is greater than 0.25).

produce a map more related to semantic and main topics. If we had morphosyntactic information, we could choose to rely on adjectives, or on verbs, as new prisms through which the corpus can be split. We could also study our corpus from the point of view of a given lexical field and its implementation throughout the texts.

4. Conclusion

4.1 Textometry's relevance for research in the Humanities

Textometry (Lebart, Salem, and Berry 1998) can be seen as a relevant approach for Humanities research, as analysis is directly grounded in texts, not necessarily mediated by external linguistic or semantic resources, which may not be fully appropriate for the corpus and thus could bias what can be seen and found. Information is, above all, taken directly from the corpus, texts, and contexts, and does not depend on available resources, which might not be precisely adjusted to every kind of data. Lexicographical resources can be used anyway, and can add valuable information, but they do not play a critical role; one does not rely on them to access and analyze the texts. The corpus can be seen, not only through an external lexicon, but also through the words it actually contains and the way they are used. This is the first important reason to be interested in textometry: texts and corpus come first, and they are a core foundation during the whole process. The corpus is the main source of information about the texts it includes, and the way those texts use words. Moreover, textometry takes the corpus as it is: one does not have to remove unknown words; one is not compelled to normalize word forms.

A second feature of textometry is that it is not an automatic approach. In a textometric analysis, the computer does not replace the researcher, it does only what it is best at: storing and processing—that's all. Of course, processing can be complex, and it often goes much further than simple counting: as shown above, textometric functions implement algorithms, score tests, and perform complex statistical functions and elaborate visualization processes.

Actually, computers are dedicated to such intensive calculations, the very kinds of tasks for which textometry uses them. Even if a computer does a great job due to its memory and processing ability, even if it allows statistical tests or analytic procedures that would otherwise be impossible, the researcher controls all aspects of the investigations, and bears ultimate responsibility for the relevance and interpretation of computed results.

While the corpus determines the context of observations and acts as a reference concerning word frequencies, its composition significantly proceeds from a human hand and a scientific

decision, not from an automatic production. This sheds light on the *quality* of the corpus (the corpus elicits interest, the researcher is familiar with it, it is carefully composed) instead of on its *quantity* (of course, a small corpus may not need any computational tool, but the hugest return of harvested data does not give any guarantee of meaningful results either).

Once the corpus has been defined, the researcher keeps on driving the analysis, finding entrance points and giving sense to the results produced by queries and statistics. Textometry is based on a strong methodology; it offers a selective choice of tools that are most relevant for textual data, but there is no unique or predefined path to get results, no unique reading of a corpus. One cannot say, “let's just try what textometry tells us about this corpus,” because there is not a given output for a given input, and even the input is a considered matter.

As shown in our examples, textometry combines quantitative and qualitative processing. Statistics or basic heuristics are defined that make sense on textual data (see section 5, technical appendix). Even qualitative KWIC concordance view is specifically designed to emphasize linguistic properties of the data. KWIC view is not just an exhaustive compilation of extracted contexts containing a given keyword. A textometric concordance combines several kinds of contextual displays: a tabulated view with context sorting, so as to reveal close contextual patterns, associated with references indicating useful upper-level metadata; and a page view in a full edition, providing all the graphical and typographical clues for reading and interpretation.

In such an analytic framework, textual editions can still meet requirements for textual studies in the Humanities, according to the researcher's experience that reading, analyzing, and editing, are not separate activities. TXM software, for example, can work on TEI encoded corpora, display fine HTML editions, and manage several aligned versions of the same text, including multimedia capture of the original source (for instance pictures of manuscripts).

4.2 Experiencing textometry on new corpora

The purpose of this paper was not primarily to establish new scientific results about Plato's *Gorgias*, but it aims to introduce the classicist scientific community to the textometric approach to digital studies. Textometry could be characterized as a both qualitative and quantitative analysis methodology. Behind a detailed typology of queries that could be searched for in the corpus (queries about word frequency, syntagmatic or paradigmatic uses of word meaning in the corpus, word evolution, texts’ contrastive characterization and corpus internal structures),

the key idea is that it is possible in a digital context to elaborate tools combining advanced computational analysis and philologic attention and sensibility towards text integrity and richness.

We hope that this approach can become a personal experience for many colleagues, since several applications implement the textometric ideas promoted here. Our example has been realized with TXM open-source software, which is available for multiple operating systems (including a web portal version), most languages (including Latin and ancient Greek), and many corpus encoding states (from raw text corpora to TEI encoded ones). It also gives all the resources to put into TXM any corpus built from the open Perseus Digital Library.

Acknowledgments

We are grateful to our two reviewers for their accurate comments and stimulating suggestions.

5. Technical Appendix: More Information about ThreeTextometric Processes

5.1 Specificities: A statistical computation to find keywords

The specificity score (Lebart, Salem, and Berry 1998) is an application of Fisher's exact test to textual data (Gries 2012). Specificity analysis allows one to identify which words are specifically overused and underused in a part of the corpus, compared to its use in the whole corpus. It is like a keyword analysis in some textual analysis programs, but instead of using a log-likelihood, t-score, z-score, or tf.idf measure, it implements Fisher's exact test, which is known as the most accurate measure for words' frequency distributions (Gries 2014; McEnery and Hardie 2012).

Based on a hypergeometric probability model, this analysis provides either a positive or negative specificity score. A *positive* score indicates the order of magnitude of the probability of a word *w* appearing *f* times *or more* than *f* times in a part containing *n* words, given that *w* appears *F* times in whole corpus of *N* words. A *negative* score is given when the frequency is lower than expected on a random basis; the measured probability is then the one for a word appearing *f* times *or less* than *f* times, given its total frequency *F*, and part and corpus sizes (*n* and *N*). Specificity scores equal to or higher than 3 (1 chance in 1,000 to obtain the frequency f or more if the words were randomly distributed among parts) or lower than -3 (1 chance in 1,000 to obtain the frequency f or less randomly) are considered significant.

Figures 6, 7, 9, and 20 show examples of specificities outputs, which point out lexical choices of some texts within Plato's work.

5.2 Correspondence analysis: A geometrical optimization to get a word-based map of the corpus

Map visualization of the corpus is based on correspondence analysis, a multivariate statistical tool. The idea behind correspondence analysis may be explained in a few words. Every text is represented according to the words it uses: geometrically speaking, each text is a point and gets a unique location in a multidimensional space in which each word is an axis. The coordinate of a text point on the *w* word axis is related to the frequency of the *w* word in the text. If two texts have many words in common and use them with similar frequencies, then their points get close locations in space. Mathematical transformations are then used to build a new set of axes for this space, with new worthwhile properties. The new axes combine the previous word-axes (each axis is a linear combination of word-axes) so that: (i) all the information is kept (points representing texts keep their distances to one another, the shape of the "text-cloud" is the same); (ii) redundancy is eliminated, which reduces the number of axes; (iii) the new axes are ordered so that the first one shows the "largest" and "heaviest" variations (i.e. along this axis frequencies are very different and concern many words and texts), and each following axis brings the "largest" and "heaviest" remaining variations, so that with axes from 1 to any number *N*, one gets the best *N*- dimensional view of the lexical variations and text oppositions inside the corpus. Given the value 2 for *N*, the corpus can then be visualized as a map using the Cartesian coordinate system, knowing that this map is mathematically proven as being the best 2D-visualization in order to focus on the greatest differences inside the corpus. What is to be read on this map is a quantitative summary of the main correspondences and oppositions among texts. The axes are complex weighted combinations of words whose composition might be studied, but the main use of correspondence analysis for text analysis is to look at the relative positions of texts, and axes are first used to provide the best angle to get the best view.

Figure 22 applies correspondence analysis to Plato's work, so that one gets a view about text similarities or oppositions for high frequency word usage.

Correspondence analysis is conceptually similar to principal component analysis, but correspondence analysis is preferred to principal component analysis here because it better fits the data. In the field of textometry, textual data are represented in frequency tables where rows are words (or other kinds of linguistic units), and columns are texts (or groups or parts of texts

dividing the corpus). In this kind of table, rows and columns are both categorical variables (a set of words and a set of texts) and play symmetric roles, so structurally this is a two-way table (also called a cross tabulation, or a contingency table). Correspondence analysis is specifically designed and relevant for the analysis of such contingency tables, for which it proves getting better results than principal component analysis (Lebart, Salem, and Berry 1998:63-69).

5.3 "Back-to-text": Concordances and word-in-context functions as main features

What is called the "back-to-text" functionality in the textometric field is actually the core of textometric processing: every result (lexical list, table, or graph) must be interpreted with an eye toward corresponding words in context.

This could be a distinctive feature of a textometric approach, an efficient way to distinguish it among many current text mining proposals. With a text mining approach, one digs into texts in order to extract pieces of information, so that all the analysis focuses on these few well identified pieces instead of an unstructured full text; and visualization is often the goal, the end of the analysis, with a synthetic and suggestive view that replaces and summarizes the corpus. On the contrary, the textometric approach chooses and implements the hermeneutic circle: any statistical summary or view has to be interpreted looking back to the text, any distant reading view invites renewed close reading, so analysis cannot get rid of textual richness and complexity—it is grounded on this textual source material.

Bibliography

Benzécri, Jean-Paul, *et al.* 1973. *L'analyse des données*. Vol. 1, *La taxinomie*. Vol. 2, *L'Analyse des Correspondances*. Paris.

———. 1981. *Pratique de l'analyse des données*. Vol. 3, *Linguistique & lexicologie*. Paris.

Brunet, Étienne. 2009. *Écrits choisis*. Vol. 1, *Comptes d'auteurs: Études statistiques de Rabelais à Gracq*. Ed. Damon Mayaffre. Paris.

———. 2011. *Écrits choisis*. Vol. 2, *Ce qui compte: Méthodes statistiques*. Ed. Céline Poudat. Paris.

———. 2016. *Écrits choisis*. Vol. 3, *Tous comptes faits: Questions linguistiques*. Ed. Bénédicte Pincemin. Paris.

Christ, Oliver. 1994. "A Modular and Flexible Architecture for an Integrated Corpus Query System." In *Papers in Computational Lexicography (Complex '94)*, 22–32. Budapest.

Evert, Stefan, and Andrew Hardie. 2011. “Twenty-First Century Corpus Workbench: Updating a Query Architecture for the New Millennium.” In *Proceedings of the Corpus Linguistics 2011 Conference*. Birmingham.

Gries, Stefan Th. 2012. “Corpus Linguistics: Quantitative Methods.” In *The Encyclopedia of Applied Linguistics*, ed. C. A. Chapelle, 1380–1385. Oxford.

———. 2014. “Quantitative Corpus Approaches to Linguistic Analysis: Seven or Eight Levels of Resolution and the Lessons They Teach Us.” In *Developments in English: Expanding Electronic Evidence*, ed. M. Kytö Taavitsainen, Cl. Claridge, and J. Smith, 29–47. Cambridge.

Heiden, Serge. 2010. “The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme.” In *24th Pacific Asia Conference on Language, Information and Computation - PACLIC24*, ed. R. Otoguro, K. Ishikawa, H. Umemoto, K. Yoshimoto, and Y. Harada, 389–398. Tokyo.

Heiden, Serge, Matthieu Decorde, and Sébastien Jacquot. 2018. *TXM User Manual: Version 0.7 Alpha*. Trans. Sara Pullin. Lyon. http://textometrie.ens-lyon.fr/files/documentation/TXM%20Manual%200.7.pdf.

Lafon, Pierre. 1984. *Dépouillements et statistiques en lexicométrie*. Paris.

Lebart, Ludovic, André Salem, and Lisette Berry. 1998. *Exploring Textual Data*. Boston.

Léon, Jacqueline, and Sylvain Loiseau, eds. 2016. *Quantitative Linguistics in France*. Lüdenscheid.

Marchand, Stéphane, and Pierre Ponchon. 2016. Gorgias *de Platon suivi de* Éloge d’Hélène *de Gorgias*. Paris.

Marchello-Nizia, Christiane, and Alexei Lavrentiev. 2013. *Queste del saint Graal: Édition numérique interactive du manuscrit de Lyon (Bibliothèque municipale, P.A. 77)*. Lyon. http://catalog.bfm-corpus.org/qgraal_cm.

McEnery, Tony, and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge, MA.

O’Sullivan, Neil. 1993. “Plato and ἡ Καλουμένη Ῥητορική.” *Mnemosyne* 46:87–89.

Salem, André. 1987. *Pratique des segments répétés: essai de statistique textuelle*. Paris.

Schiappa, Edward. 1990. “*Did Plato Coin Rhètorikè.*” *The American Journal of Philology* 111:457–470. https://doi.org/10.2307/295241.

———. 1999. *The Beginnings of Rhetorical Theory in Classical Greece*. New Haven and London.