New digital tools for the study of Greek papyri

Currently developed by the Grupo de Lingüística Griega (ILC, CCHS, CSIC)

Callimachus

Callimachus is an automated regest of published papyri and ostraka, ie. a processed extract of the formal contents of the text in the papyri hosted at the Papyri.info site. Additional information about the date, origin, material, etc., of the papyri (from the HGV database) is included in order to enrich the queries. Callimachus contains data on both documentary and literary papyri. The *lexical* information about the papyri is contained in the sibling *Polyphemus* database 👉.

Callimachus contains three kinds of information:

❧ The first type of information refers to *several countable features of the text,* as it was encoded by the Papyri.info project. For example, how many words, letters, gaps, letters per line, scribal hands, etc. can be found inside every document. This data was extracted during the parsing of the documents from the Integrating Digital Papyrology Papyri.info Github repository.

❧ The second type of information is an *automated calculation of the state of the text of the papyrus* ***(Callimachus' number).*** In other words, how much (and how well) the original text of the papyrus can be read in the edition used by Papyri.info. This calculation is provided as two decimal numbers (CRN and CNN) from 0 to 1 (one means all the text is perfectly readable).

> *Callimachus Readability Number (CRN) is a measure of the readability of the part of the text that was edited (up to which point the editor was able to read or conjecture the papyrus' text information).*
>
> *Callimachus Conservation Number (CCN) is a measure of the ste of conservation of the papyrus' text.*

❧ The third kind of data is mainly *data about the papyrus (or ostrakon) itself, as provided by the Papyri.info project:* date, origin, material, content, etc. This information comes from the metadata included in the XML documents, or from the HGV database.

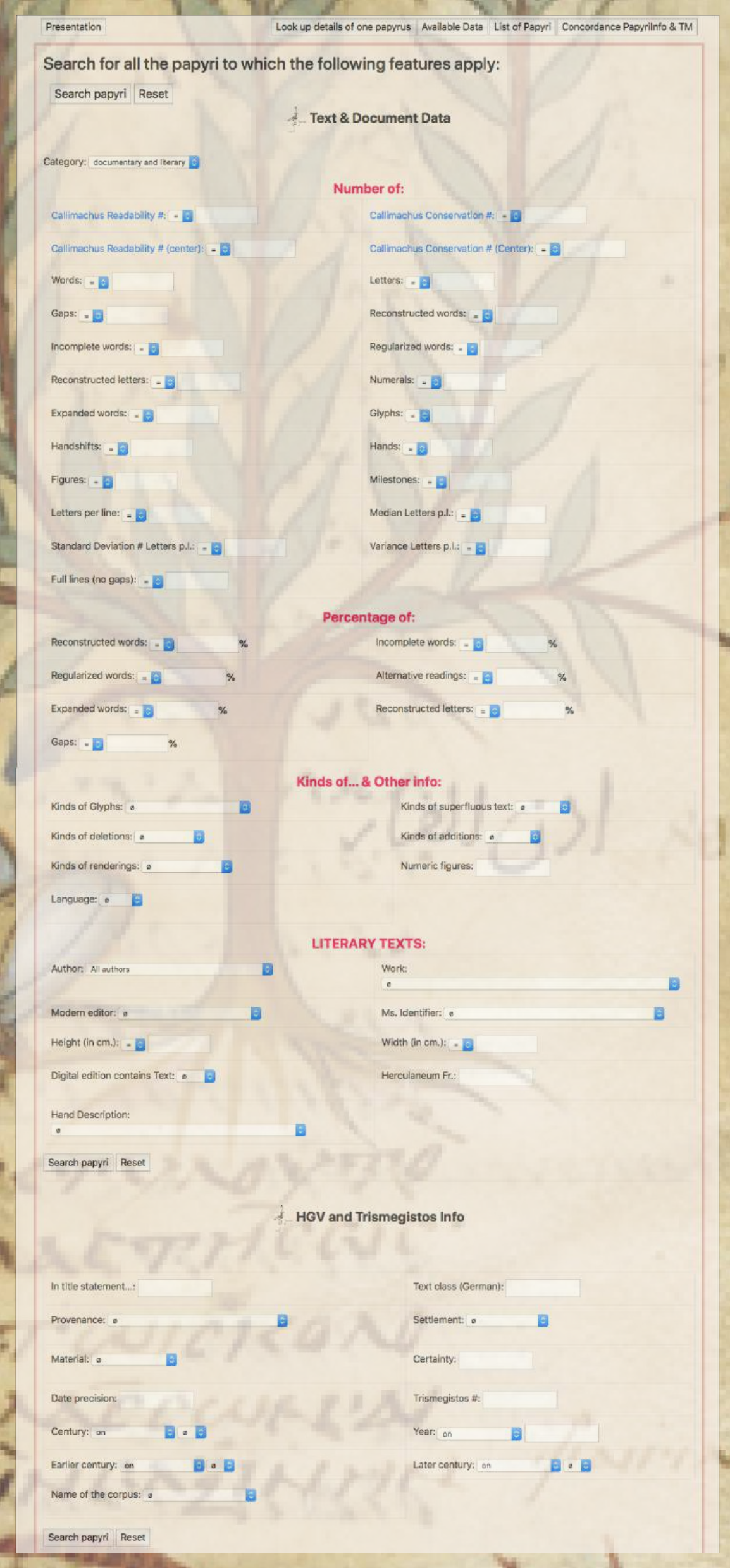
What is Callimachus for?

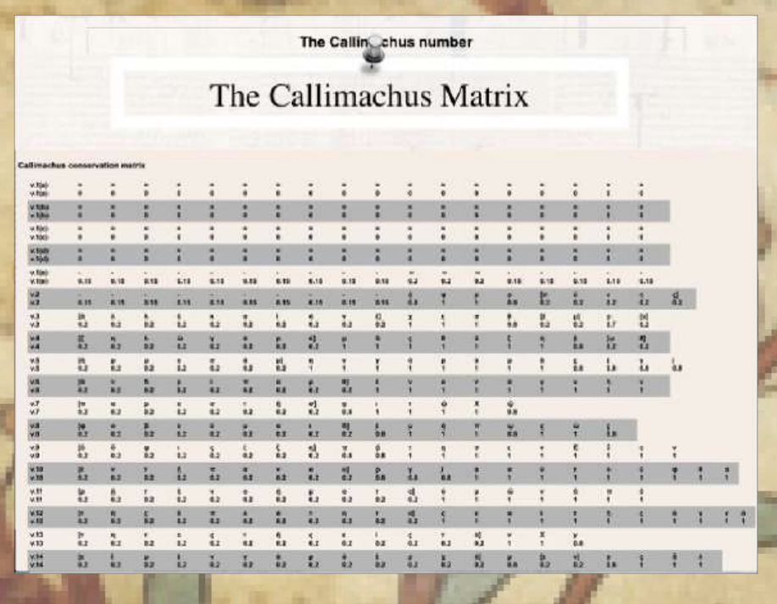
Callimachus can be used in papyrology research as well as in linguistics-related projects.

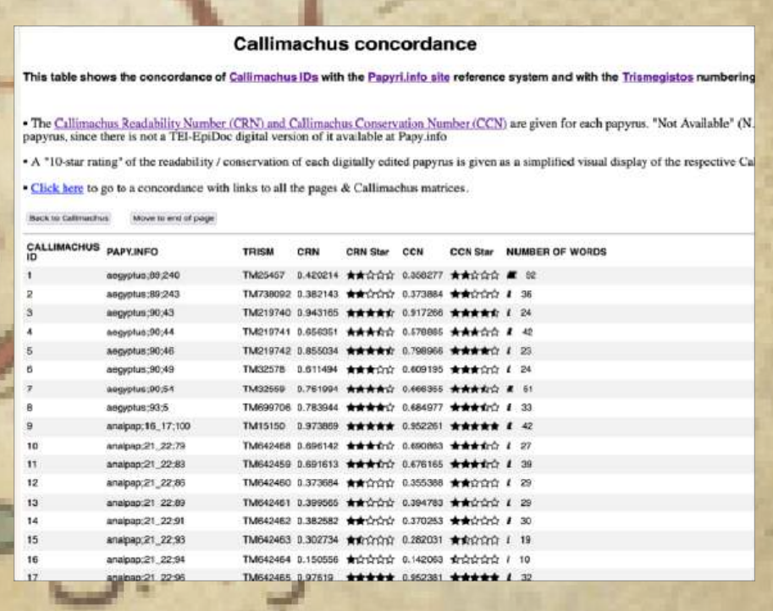
You can use Callimachus to search papyri containing any specific feature, or a combination of features. For example, you can search for papyri containing any specific trait (Are there coronides in non literary papyri? Where can I find examples of papyri using a specific fraction, or a type of deletion mark?, etc.), or combination of traits. This may help you to find parallels to your object of study. The use of Callimachus in combination with *Polyphemus*👉 will allow you to combine lexical information with all the data types of Callimachus.

You can use Callimachus to help you build your corpus, using, for example, mainly papyri from a certain date and origin with high Callimachus number (meaning better preservation) and with a minimum of words.

Some features (like a high number of regularizations) can be meaningful for the linguist interested in phonetic traits on the koiné.


Search for all the papyri to which the following features apply:
Text & Document Data
Number of:
Percentage of:
Kinds of... & Other info:
LITERARY TEXTS:
HGV and Trismegistos info





Polyphemus

Polyphemus is *a database of morphologically annotated Greek papyri, ostraka, graffiti, etc.* (hence we'll use just "papyri"). The digital version of the papyri was originally encoded by the partners of the papyri.info project using EpiDoc, an adaptation of the TEI scheme for ancient documents.

All texts from the papyri.info site are parsed by the Polyphemus engine. Actual words that were read or philologically recovered by the editors are differentiated from letters that can not be made part of a word. We try to identify the lemma, and assign a part-of-speech (POS) tag and morphological parsing to each word. More than 98% of the actual words can by lemmatized and assigned a full grammatical tag. All lemmatized words are assigned a gloss by the *Madrid Ancient Greek WordList* 👇.

Since the papyri.info collection of documents is in a permanent state of revision and growth, we try to update our database on a fairly regular basis (at least four times a year).

What is Polyphemus for?

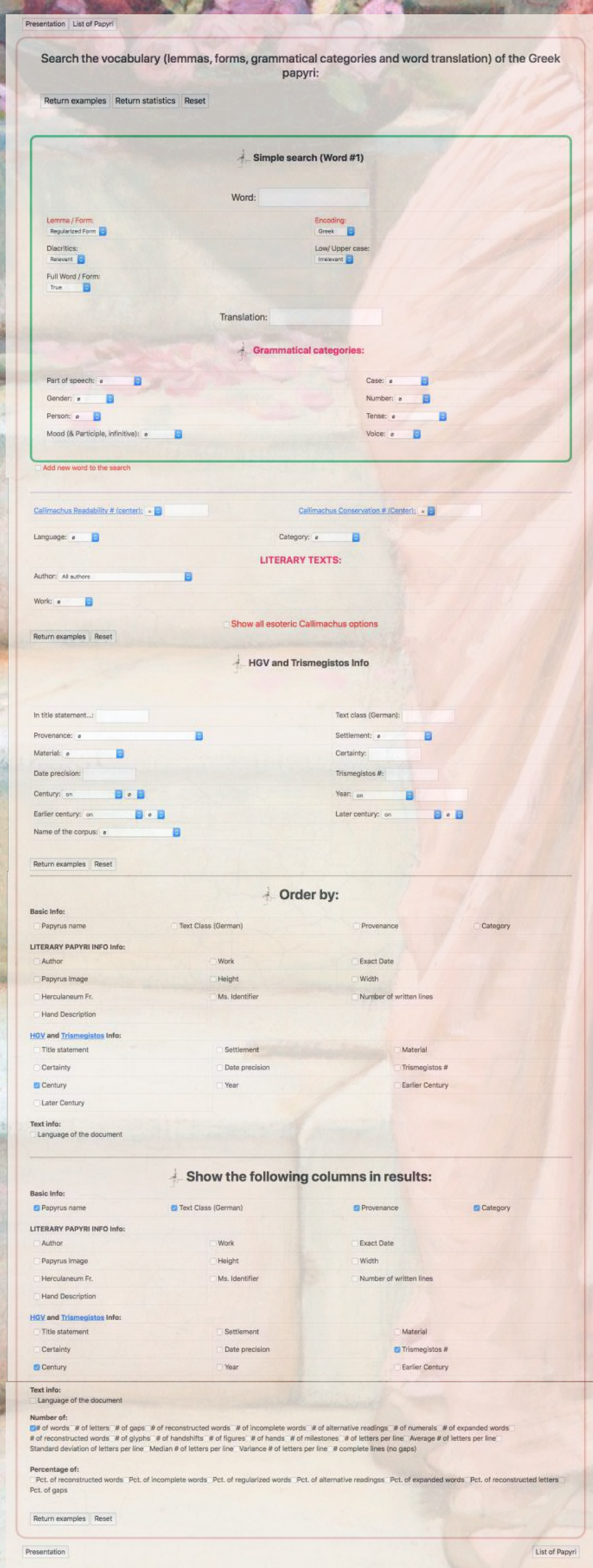
Polyphemus can be used in papyrology or linguistically oriented research, as well as in historical, sociological or anthropological projects. It has a clear pedagogical side apt for educational projects that aim to discover the social, economical, intellectual and even personal side of the people who wrote and read these ancient texts.

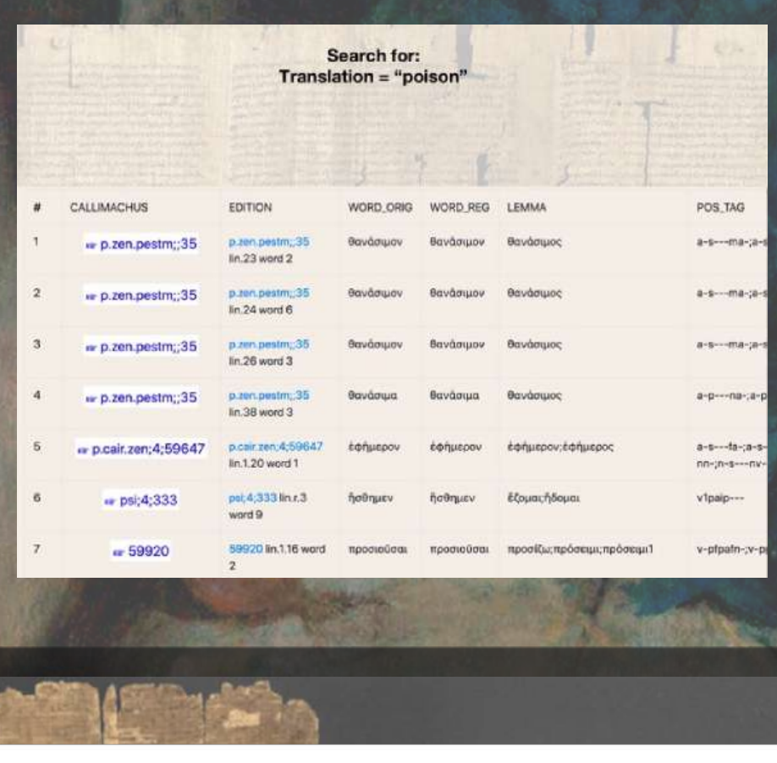
You can use Polyphemus to search for any word, combination of words, translation or grammatical form present in any papyri contained in the papyri.info database. You can combine any linguistic or grammatical parameter (lemma, POS-tag, translation) with any non grammatical parameter used by the 👈*Callimachus* database.

Soon you will be able to request from Polyphemus statistics on any query that returns many results. These statistics provide information on chronological or geographical variation with statistical weighted results.

Practical examples

You can use Polyphemus to know the exact place on the papyri of any given century, or location, where any word appears. You can use it in order to know where any word meaning "poison" or "fruit", or "slave", etc., appears in combination with any other word, on any location, in documents made on any material…


Search the vocabulary (lemmas, forms, grammatical categories and word translation) of the Greek papyri
Return examples
Reset
Simple search (Word #1)
Grammatical categories
LITERARY TEXTS:
HGV and Trismegistos info
Order by:
Show the following columns in results:
Presentation
List of Papyri



Treebank of the *Papyri*

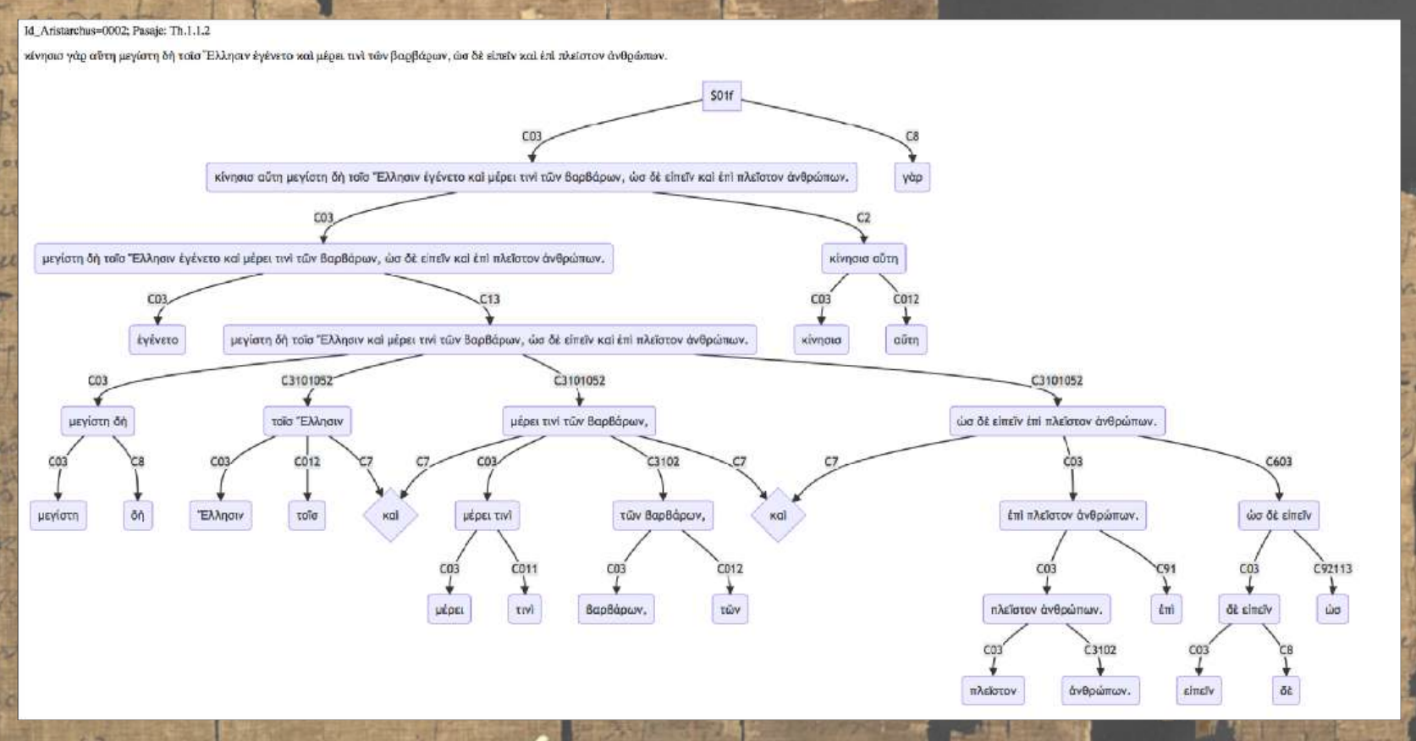
The aim of this project is to create a treebank of all the texts of Philodemus of Gadara (an epicurean Greek writer of the first century BC) found in the Villa dei Papiri, in Herculaneum. In time, our treebank will cover all the edited papyri form Herculaneum. To do this, we use *MAGWL* 👇 to provisionally annotate every word, and then we proceed to manually annotate syntactically and semantically all the texts, using the computer editor Aristarchus (Riaño 2006).

This annotation includes syntactic functions, some semantic functions, anaphora, and morphological composition. It also records the precise preservation status of each letter of the text when it was edited (whether the text is/ was clearly readable in the document, if the lecture is based on the remains of the letter, or if it is the editor's conjecture, etc.) The syntactic annotation is carried out using Immediate Constituent analysis. Finally we transform the text analysed employing such schema into two different treebanks: one that uses a Immediate Constituent description, and another one that results of a transformation into the Dependency Analysis principles using a Stratified Dependency Analysis description, closely following PROIEL schema, compatible with the Universal Dependencies guidelines. (The syntactic tree👉 belongs to Thucydides.1.1.2).

How do we transform an Immediate Constituent treebank into a Dependency treebank?

A Constituent treebank can be partially converted into a dependency treebank using a "stratified" approach:

A. Immediate constituents have *stratum* number.
B. Every word depending on the same head belongs to a different *stratum*.
C. An IICC is annotated with a syntactic tag whenever its function is not just being the nucleus of the syntagm
D. If we call right branch the second branch of every IICC, the *stratum* number is the number of right branches from the higher constituent where a word appears as the nucleus of the syntagm to the lower constituent where its parent is the nucleus.



```
<sentence ID_oracion="Th.1.1.1" ID_doc="0001"><c id="$01f" words="001.002.003a.
004.005.006.007.008.009.010.011.012.013.014.015.016.017.018.019.020.021.022.023.024.025.026.027.028.029.030.031.032.033.034.035.036.037.038.039.040.041.042.043.044.045.046.047.048.049.050.051.9
001.9002" nuc="003" relation="PRED" codari="C03" IC="1">Θουκυδίδησ Ἀθηναῖοσ ξυνέγραψε τὸν πόλεμον τῶν Πελοποννησίων καὶ Ἀθηναίων, ὡσ ἐπολέμησαν πρὸσ ἀλλήλουσ, ἀρξάμενοσ εὐθὺσ
καθισταμένου καὶ ἐλπίσασ μέγαν τε ἔσεσθαι καὶ ἀξιολογώτατον τῶν προγεγενημένων, τεκμαιρόμενοσ ὅτι ἀκμάζοντέσ τε ἦσαν ἐσ αὐτὸν ἀμφότεροι παρασκευῇ τῇ πάσῃ καὶ τὸ ἄλλο Ἑλληνικὸν
ὁρῶν ξυνιστάμενον πρὸσ ἑκατέρουσ, τὸ μὲν εὐθύσ, τὸ δὲ καὶ διανοούμενον. elipt.9001.ξυνιστάμενον elipt.9002.ξυνιστάμενον
  <c id="$01f,01f" words="003a.
004.005.006.007.008.009.010.011.012.013.014.015.016.017.018.019.020.021.022.023.024.025.026.027.028.029.030.031.032.033.034.035.036.037.038.039.040.041.042.043.044.045.046.047.048.049.050.051.9
001.9002" nuc="003" relation="PRED" codari="C03" IC="1.1">ξυνέγραψε τὸν πόλεμον τῶν Πελοποννησίων καὶ Ἀθηναίων, ὡσ ἐπολέμησαν πρὸσ ἀλλήλουσ, ἀρξάμενοσ εὐθὺσ καθισταμένου καὶ
ἐλπίσασ μέγαν τε ἔσεσθαι καὶ ἀξιολογώτατον τῶν προγεγενημένων, τεκμαιρόμενοσ ὅτι ἀκμάζοντέσ τε ἦσαν ἐσ αὐτὸν ἀμφότεροι παρασκευῇ τῇ πάσῃ καὶ τὸ ἄλλο Ἑλληνικὸν ὁρῶν ξυνιστάμενον
πρὸσ ἑκατέρουσ, τὸ μὲν εὐθύσ, τὸ δὲ καὶ διανοούμενον. elipt.9001.ξυνιστάμενον elipt.9002.ξυνιστάμενον
    <c id="$01f,01f,01f" words="003a.004.005.006.007.008.009.010.011.012.013" nuc="003" relation="PRED" codari="C03" IC="1.1.1">ξυνέγραψε τὸν πόλεμον τῶν Πελοποννησίων καὶ Ἀθηναίων, ὡσ
ἐπολέμησαν πρὸσ ἀλλήλουσ,
     <c id="$01f,01f,01f,01f" words="003a" nuc="003" relation="PRED" codari="C03" IC="1.1.1.1">ξυνέγραψε</c>
     <c id="$01f,01f,01f,02" words="004.005a.006.007.008.009.010.011.012.013" nuc="005" relation="OBJ" codari="C112270504" IC="1.1.1.2" lemma="πόλεμος" form="πόλεμον" postag="N-S---MA-" Primer
Complemento="true" ="true" Referencial ó específico="true" Func. sem. Creación intelectual="true" Clase de relación="Rel. por dep." Definido="true" Era Ppe.="Era Ppe. Act. Mental Activa"
Contexto="Narración">τὸν πόλεμον τῶν Πελοποννησίων καὶ Ἀθηναίων, ὡσ ἐπολέμησαν πρὸσ ἀλλήλουσ,
```

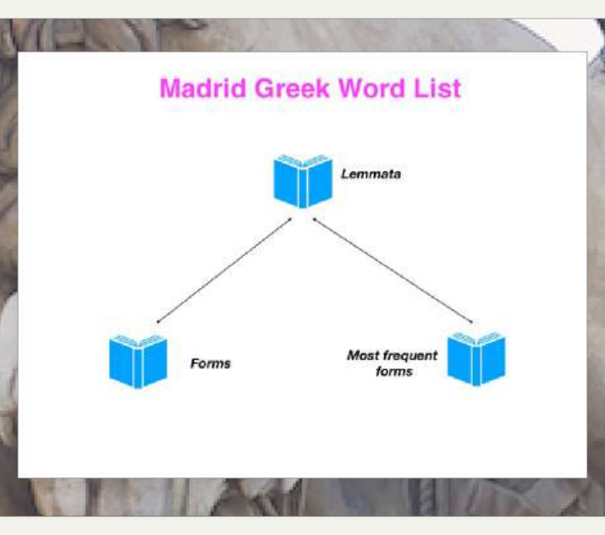
Madrid Ancient Greek Wordlist (MAGWL)

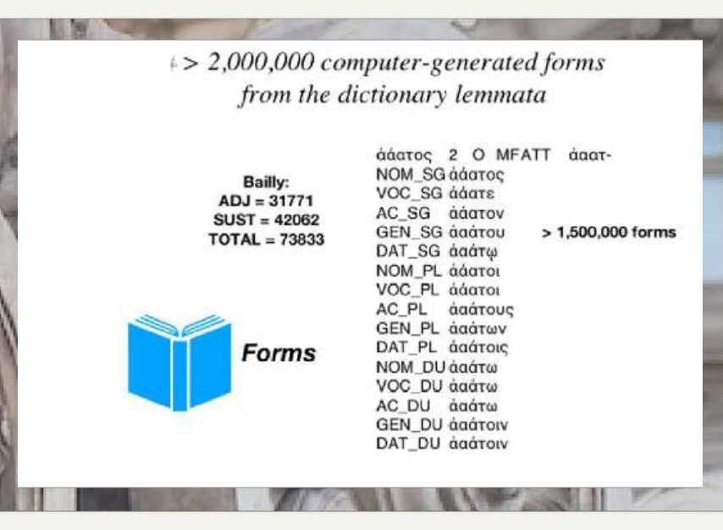
The Madrid Ancient Greek WordList (MAGWL) is a list of Ancient Greek words lemmatized, morphologically analysed, and glossed. It is mainly used to lemmatise and assign part-of-speech (POS) tags, translation, etymological& word formation data, etc., to each word-form of a text or a list. Secondarily, it can be used in all kind of projects related with the Ancient Greek world, scholarly or educational, as well as in linguistic and all kind of Digital Humanities projects.

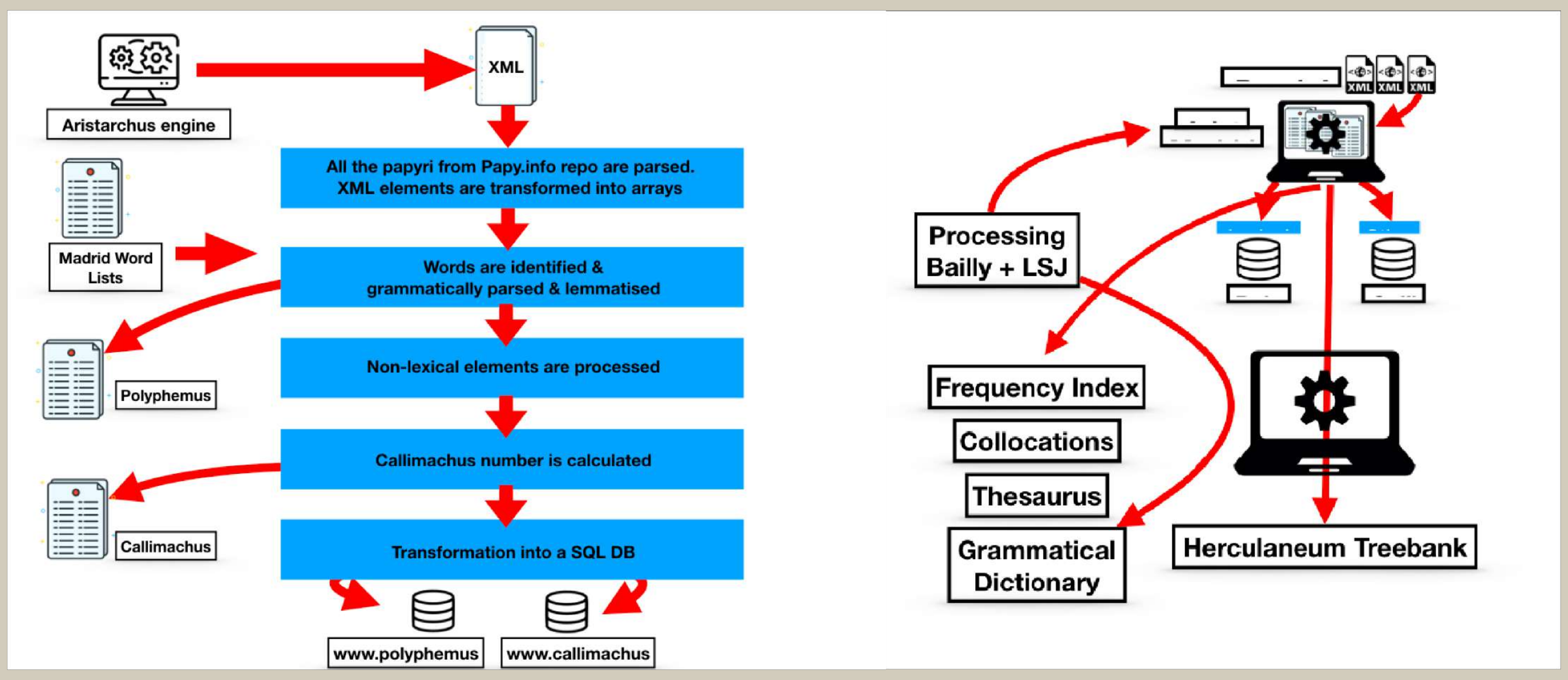
The list contains more than 1,500,000 different forms of ancient Greek words (and variant spellings for words found in papyri) and about 4,000,000 different possible morphological analysis for them. The words come from every kind of literary texts, in prose or verse, from Homer (c. 8th century BC) to the 6th century AD.

MGWL are actually three interconnected lists:

1. A list of 250,000+ lemmata (incl. proper names) with translation (glosses); citation-form, POS tagging, etymology & word formation info.
2. An ever expanding list of POS-tagged lemmatised word forms.
3. A list of the c. 40,000 most frequent forms in Ancient Greek texts.






Aristarchus engine
XML
All the papyri from Papy.info repo are parsed. XML elements are transformed into arrays
Madrid Word Lists
Words are identified & grammatically parsed & lemmatised
Polyphemus
Non-lexical elements are processed
Callimachus number is calculated
Callimachus
Transformation into a SQL DB
www.polyphemus
www.callimachus
XML
XML
XML
Processing Bailly + LSJ
Frequency Index
Collocations
Thesaurus
Grammatical Dictionary
Herculaneum Treebank

Grupo de Lingüística Griega carries out its research within the **ILC (CCHS, CSIC)** in various fields of *Greek Linguistics*: phonetics, morphology, syntax, semantics, pragmatics, stylistics and stylometry, lexicography, etymology, etc.

Although we focus on Ancient Greek (from its origins up to the 6th century AD), we do not neglect later testimonies that may shed light on the different phases of evolution of the best and most extensively documented language among those spoken today.
We consider of special relevance the combined study of literary and documentary sources, and the use of computational methods of exploitation and consultation of Ancient Greek corpora and databases. We are also actively involved in projects for the creation and development of such textual resources.

The permanent members of the Grupo de Lingüística Griega are:

❧ José Antonio Berenguer Sánchez (Investigador Científico de OPIS).
❧ Daniel Riaño Rufilanchas (Científico Titular).

Our research collaborators and predoctoral researchers include:

- ❧ Dra. Carmen García Bueno: Máster en Letras digitales (UCM) (2021-2022).
- ❧ Hugo Martín Isabel: Becario JAE Doc. (2020-2021).
- ❧ Claudia Daniela Vega Medeiros: Becaria JAE Doc. (2020-2021).
- ❧ María Belén Boned: Becaria JAE Doc. (2019-2020).

The databases and the code used to create & publish them and other projects of the *GLG* have been supported by projects FFI2017-89110-P, PGC2018-096171-B-C21 (MCIN/ AEI/10.13039/501100011033) and PID2020-118094GB-I00 (MCIN/AEI/ 10.13039/ 501100011033) from Spain's Ministerio de Ciencia, Innovación y Universidades.

GRUPO DE LINGÜÍSTICA GRIEGA, ILC, CCHS, CSIC <https://glg.csic.es>
III Workshop INTELE. Madrid 13-14 de septiembre, 2022