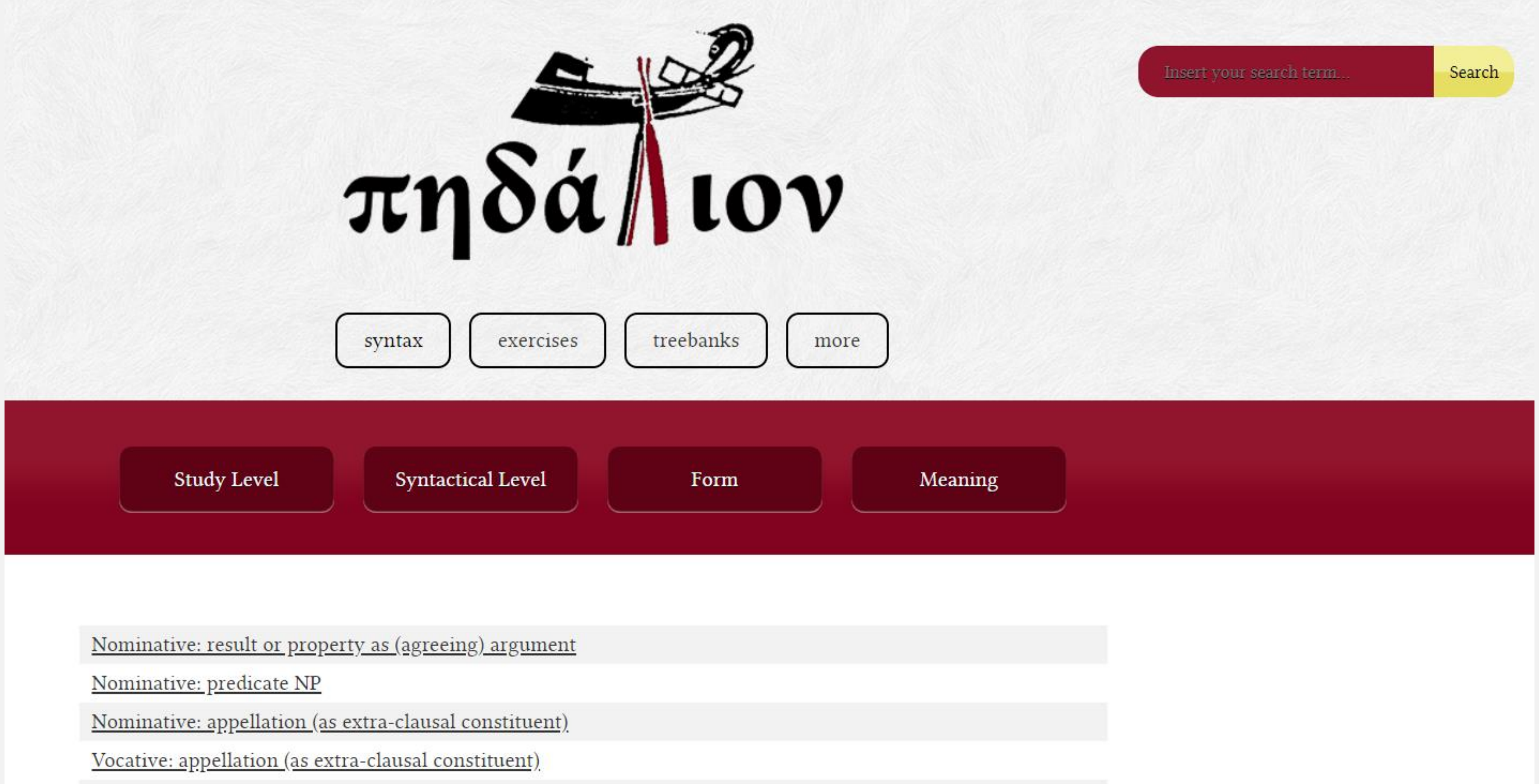
CREATING, ENRICHING AND VALORIZING TREEBANKS OF ANCIENT GREEK: THE ONGOING PEDALION PROJECT

Alek Keersmaekers (KU Leuven/Research Foundation Flanders)
Wouter Mercelis (KU Leuven)
Colin Swaelens (KU Leuven)
Toon Van Hal (KU Leuven)

PEDALION: THE ORIGINAL CONCEPT



Nominative: result or property as (agreeing) argument

Nominative: predicate NP

Nominative: appellation (as extra-clausal constituent)

Vocative: appellation (as extra-clausal constituent)

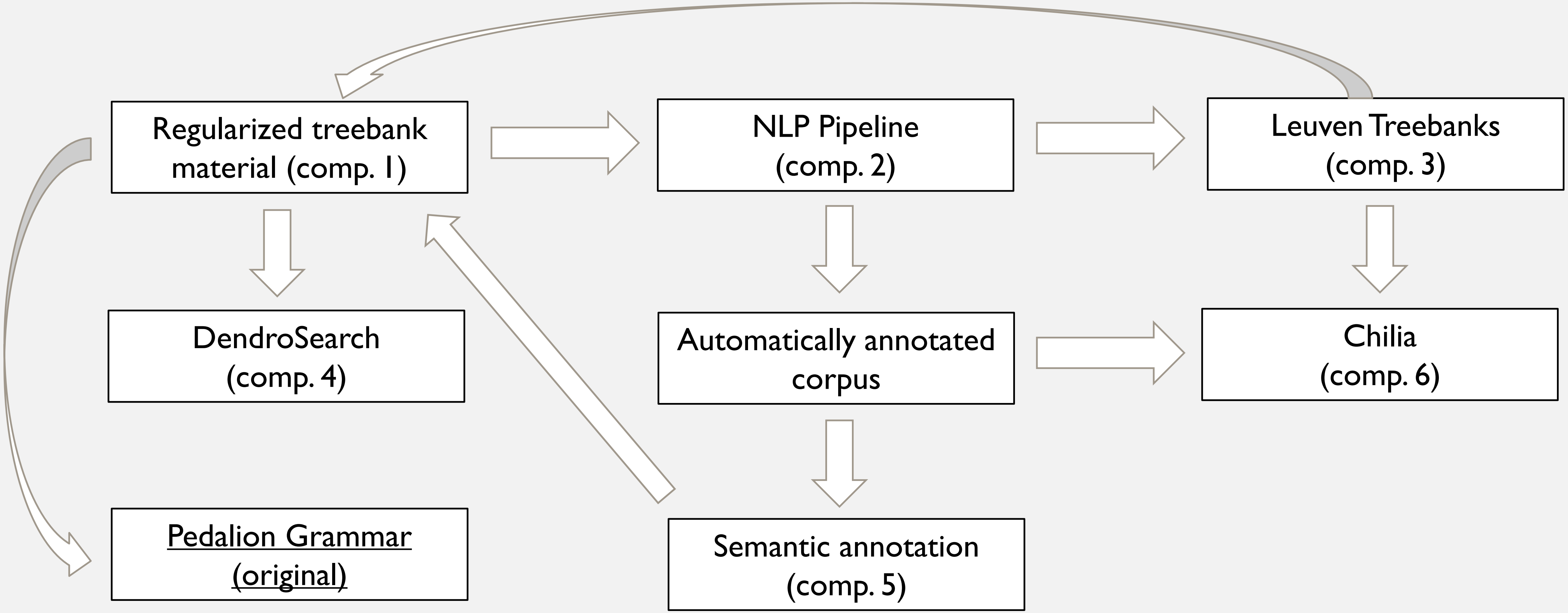
PEDALION: NEW DEVELOPMENTS

- What started as an online ‘modular’ grammar has become much larger:
  - Ancient Greek treebanks and a ‘pipeline’ to automatically create them
  - A database containing all available Greek treebank material, as well as corrections of them
  - A query tool for treebanks
  - Semantic datasets
  - The *Chilia* lexicon
- … thanks to the development of several resources for Ancient Greek

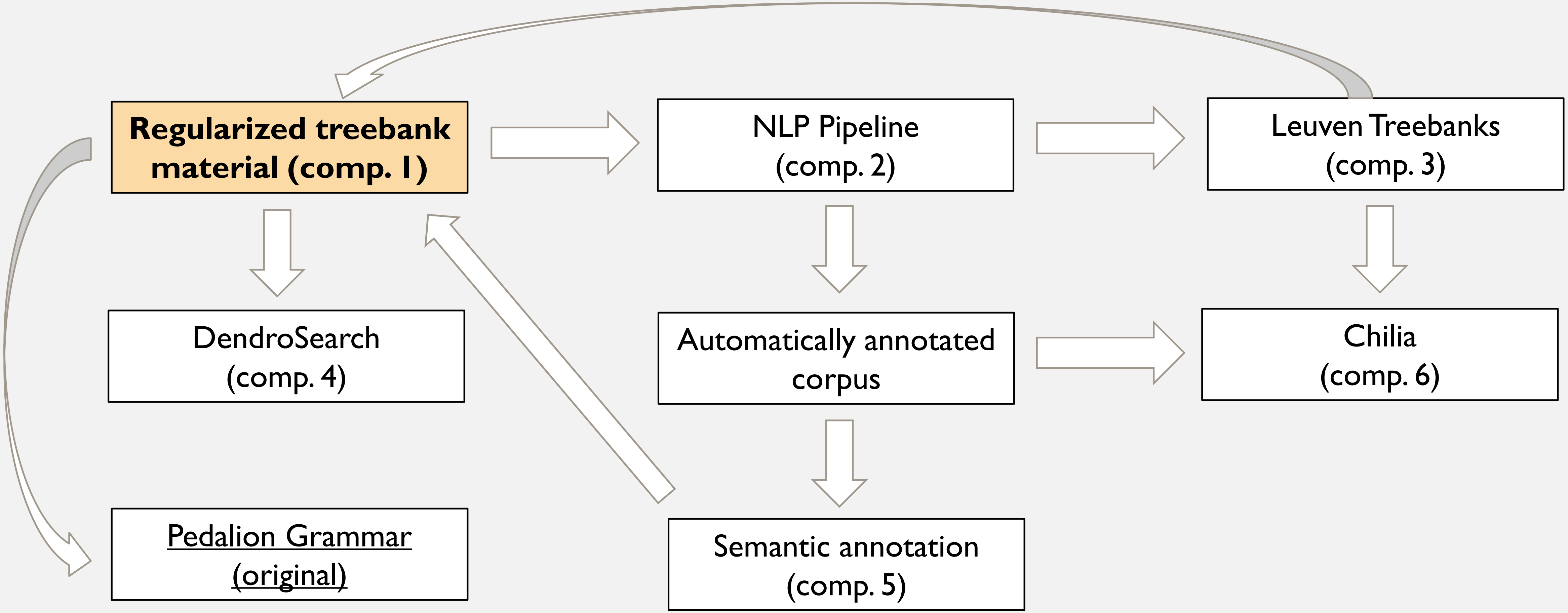
RESOURCES USED

- Integration of several existing treebanks (about 1.1 million tokens):
  - **Ancient Greek Dependency Treebanks:** 523350 tokens
  - **PROIEL Treebanks (automatic conversion):** 276433 tokens
  - **Gorman annotated trees (mostly historical texts):** 240809 tokens
  - **Harrington trees (semi-automatic conversion):** 17104 tokens
  - **Sematia/PapyGreek Treebanks (papyri):** 13017 tokens
  - **Smaller projects taken from the net:** 11158 tokens
- Other resources: e.g. Ancient Greek WordNet, PROIEL animacy dictionary, LSJ XML, Morpheus morphological analyzer, Arethusa annotator, NLP technology
- Automatically parsed papyrus and literary texts (see later)

THE PEDALION ECOSYSTEM


Regularized treebank material (comp. 1)
NLP Pipeline (comp. 2)
Leuven Treebanks (comp. 3)
DendroSearch (comp. 4)
Automatically annotated corpus
Chilia (comp. 6)
Pedalion Grammar (original)
Semantic annotation (comp. 5)

THE PEDALION ECOSYSTEM



Regularized treebank material (comp. 1)

NLP Pipeline (comp. 2)

Leuven Treebanks (comp. 3)

DendroSearch (comp. 4)

Automatically annotated corpus

Chilia (comp. 6)

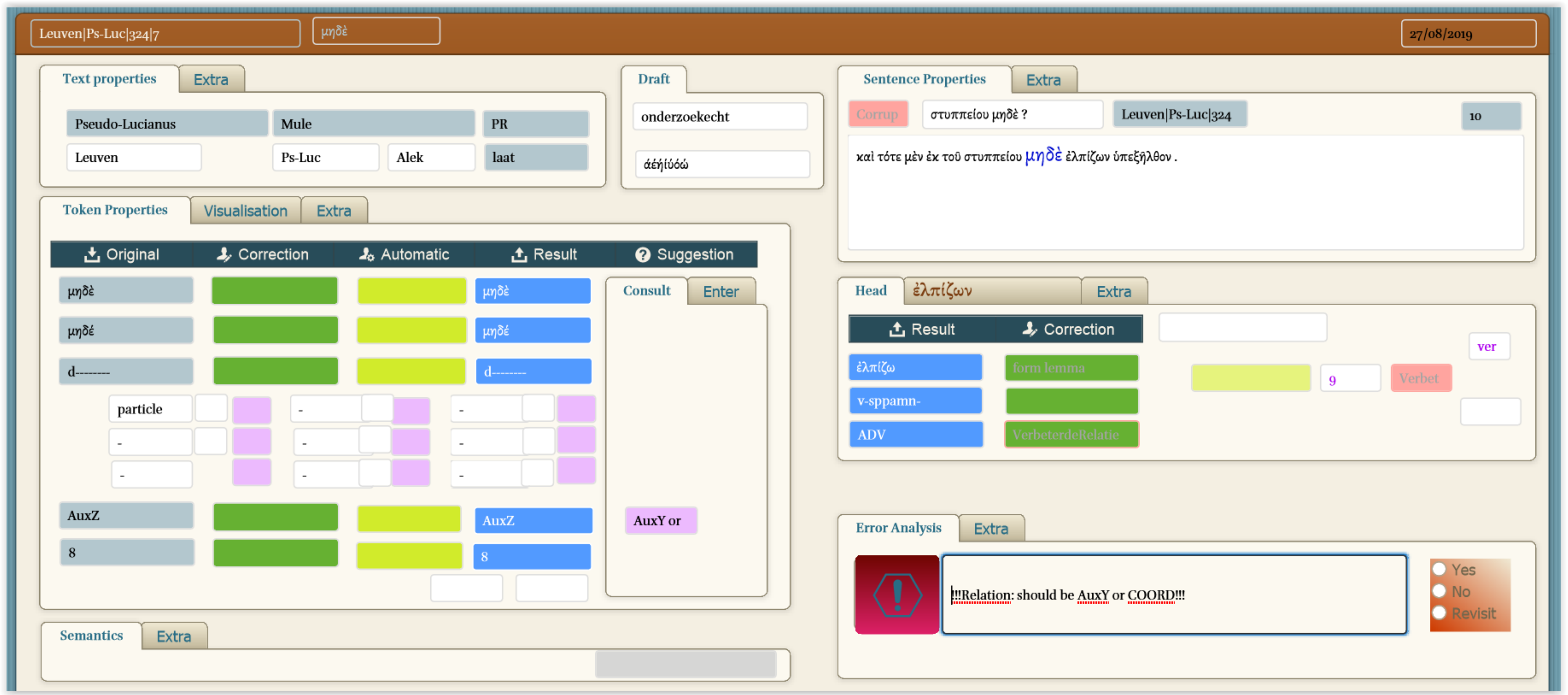
Pedalion Grammar (original)

Semantic annotation (comp. 5)

DELIVERABLE 1: REGULARIZATIONS (1)

- This project makes use of several existing corpora of Ancient Greek, each with their own differences in the annotation of specific Greek constructions
- As a result, there are a lot of inconsistencies (even sometimes in the same text from the same annotator!)
- Consistency important for NLP tasks as well as corpus linguistic research
- Therefore we integrated all these treebanks into a database (FileMaker) and are systematizing the data as much as possible
- >> https://github.com/pedalion/treebanks

DELIVERABLE 1: REGULARIZATIONS (2)



Leuven|Ps-Luc|324|7 μηδὲ 27/08/2019

Text properties | Extra

Pseudo-Lucianus	Mule	PR
Leuven	Ps-Luc Alek	laat

Draft

onderzoekecht

άέήίύόώ

Sentence Properties | Extra

Corrup στυππείου μηδὲ ? Leuven|Ps-Luc|324 10

καὶ τότε μὲν ἐκ τοῦ στυππείου μηδὲ ἐλπίζων ὑπεξῆλθον .

Token Properties | Visualisation | Extra

Original	Correction	Automatic	Result	Suggestion
μηδὲ			μηδὲ	Consult | Enter
μηδέ			μηδέ	
d--------			d--------	
particle		-	-	
-		-	-	
-		-	-	
AuxZ			AuxZ	AuxY or
8			8	

Semantics | Extra

Head ἐλπίζων Extra

Result	Correction
ἐλπίζω	form lemma
v-sppamn-	
ADV	VerbeterdeRelatie

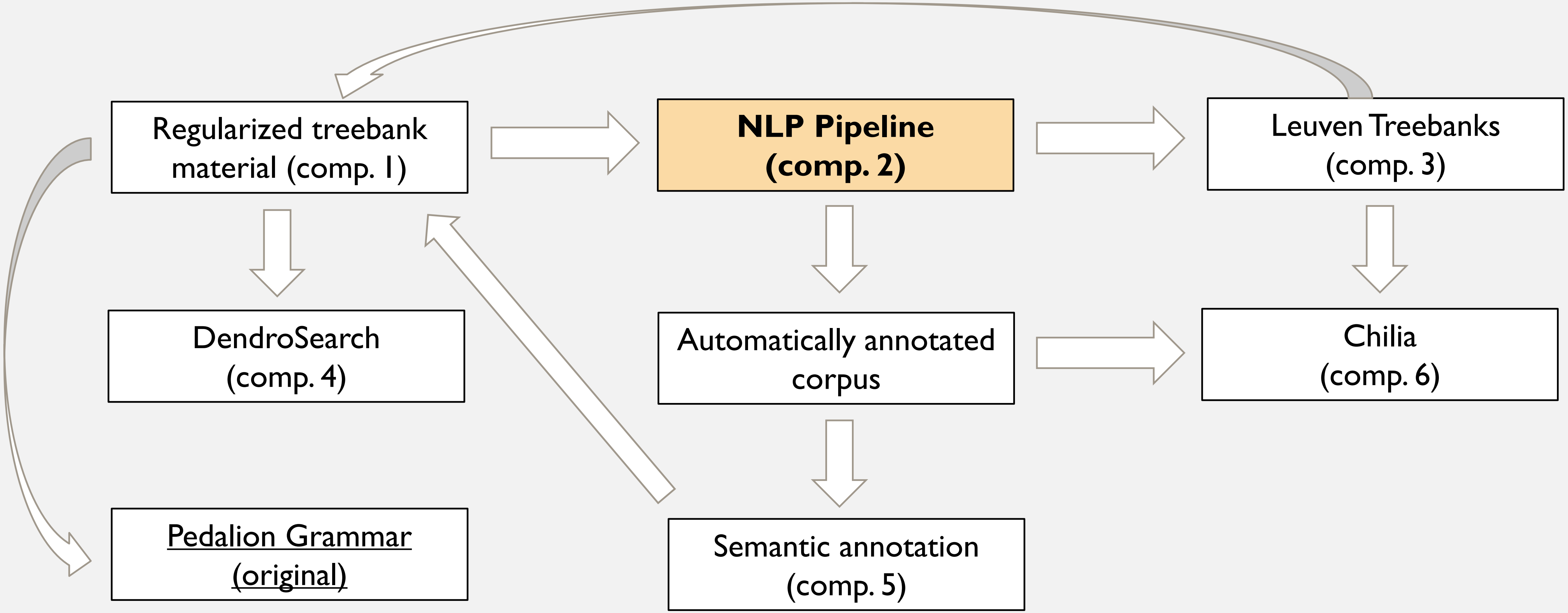
9 Verbet ver

Error Analysis | Extra

!!!Relation: should be AuxY or COORD!!!

○ Yes
○ No
○ Revisit

THE PEDALION ECOSYSTEM



Regularized treebank material (comp. 1)

NLP Pipeline (comp. 2)

Leuven Treebanks (comp. 3)

DendroSearch (comp. 4)

Automatically annotated corpus

Chilia (comp. 6)

Pedalion Grammar (original)

Semantic annotation (comp. 5)

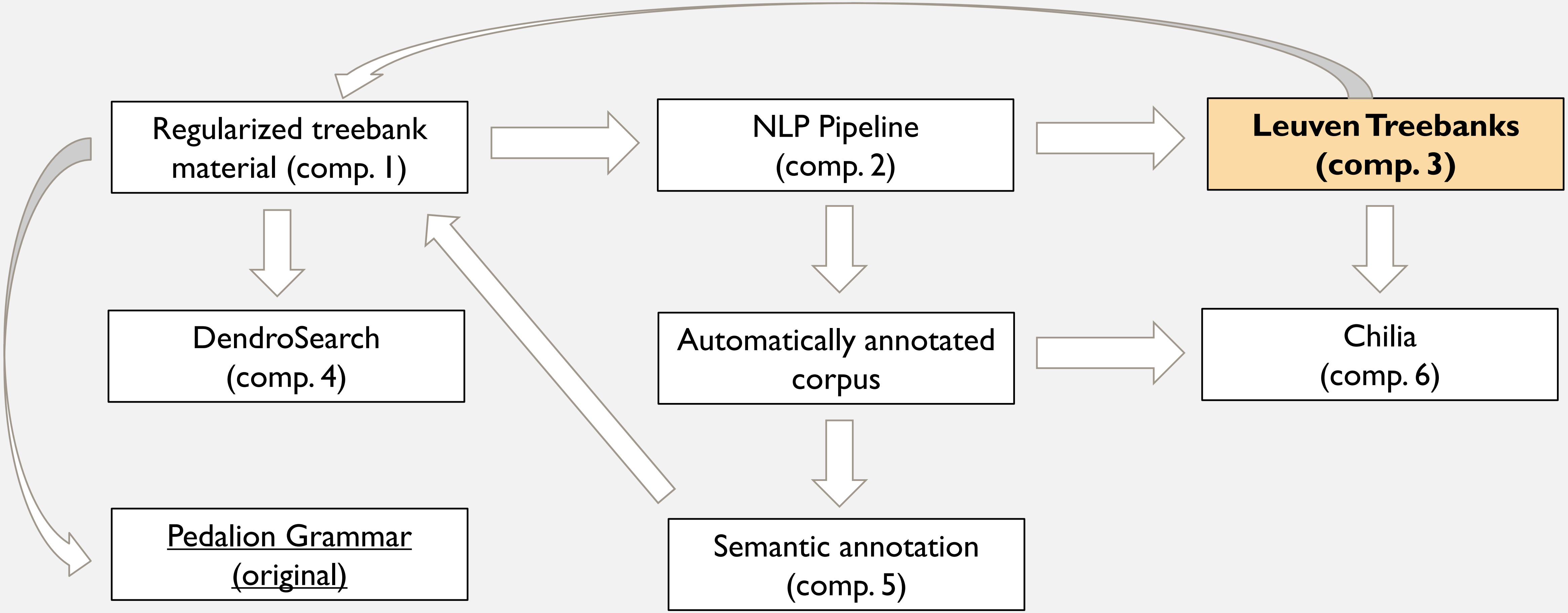
DELIVERABLE 2: TEXT ANALYSIS (1)

- These treebanks were in turn used as training material for a NLP pipeline
- Tokenization (rule-based), part-of-speech and morphology (RFTagger), lemmatization (MarMoT), syntactic parsing (Turku Neural Parser)
- Accuracy of morphological tagging is about 90% (at worst) to 96% (at best), of lemmatization 96% (at worst) to 99.5% (at best), syntactic parsing somewhere between 80-90% (difficult to assess)
- See Keersmaekers 2019, Mercelis 2019

DELIVERABLE 2: TEXT ANALYSIS (2)

- With this pipeline, we were able to automatically analyze
  - The Greek literary corpus (about 32 million tokens)
  - The papyrus corpus (about 4.5 million tokens)
- This ‘raw material’ has been used in a number of our other projects (word vectors, Chilia)
- Parts of this material has also been manually corrected

THE PEDALION ECOSYSTEM



Regularized treebank material (comp. 1)

NLP Pipeline (comp. 2)

Leuven Treebanks (comp. 3)

DendroSearch (comp. 4)

Automatically annotated corpus

Chilia (comp. 6)

Pedalion Grammar (original)

Semantic annotation (comp. 5)

DELIVERABLE 3: LEUVEN TREEBANKS (1)

- All (except for one) first automatically analyzed and then manually corrected
- Ca 160K tokens:
  - Narrative texts (35K)
  - Philosophical and scientific texts (25K)
  - Non-literary and literary letters (18K)
  - History and religious history (21K)

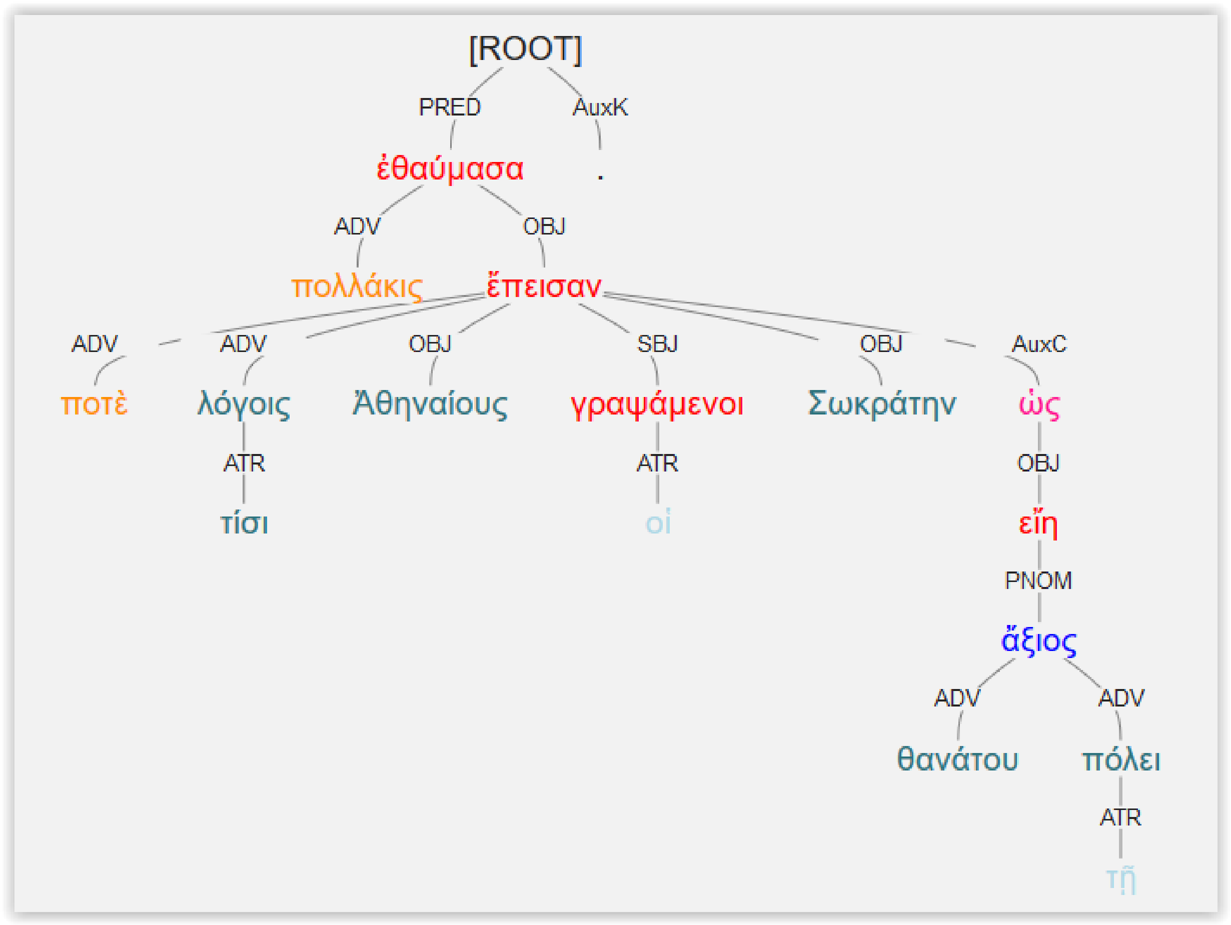
  Drama & Comedy (26K)

  Poetry (3K)

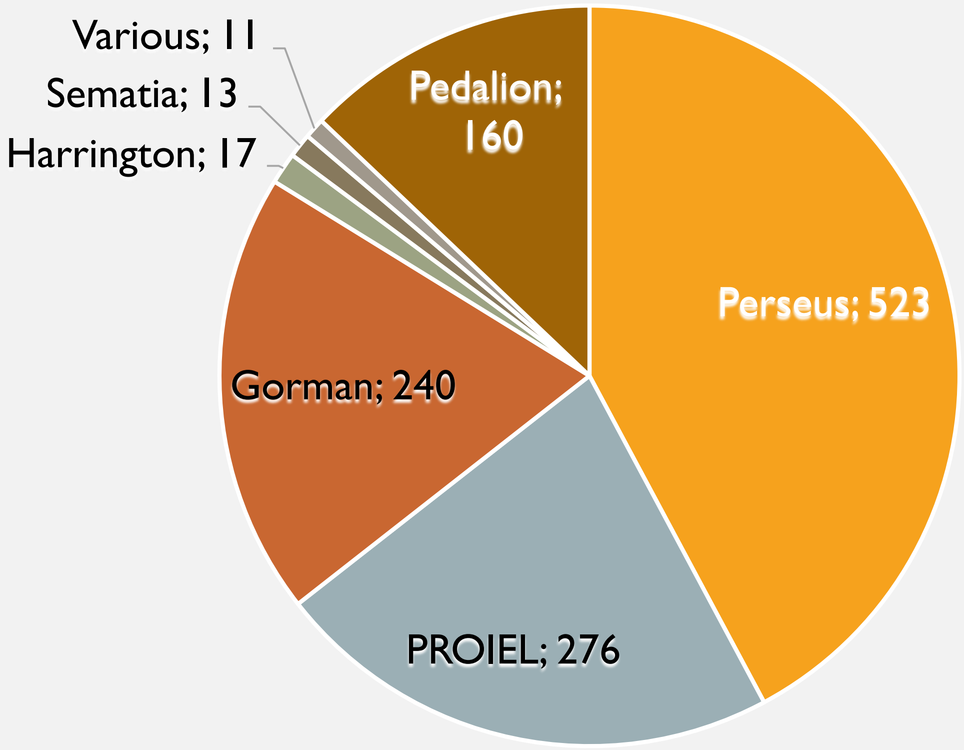
  Oratory (3K)

  Various (Chilia/Pedalion) (20K)
- Some authors are well-known (Euripides,Aristophanes, Xenophon, Lucian, Plato), other texts are less obvious (Procopius, Paeanius, Septuagint, …)
- Annotated with Arethusa by ourselves (small part), job and thesis students
- >> https://github.com/perseids-publications/pedalion-trees

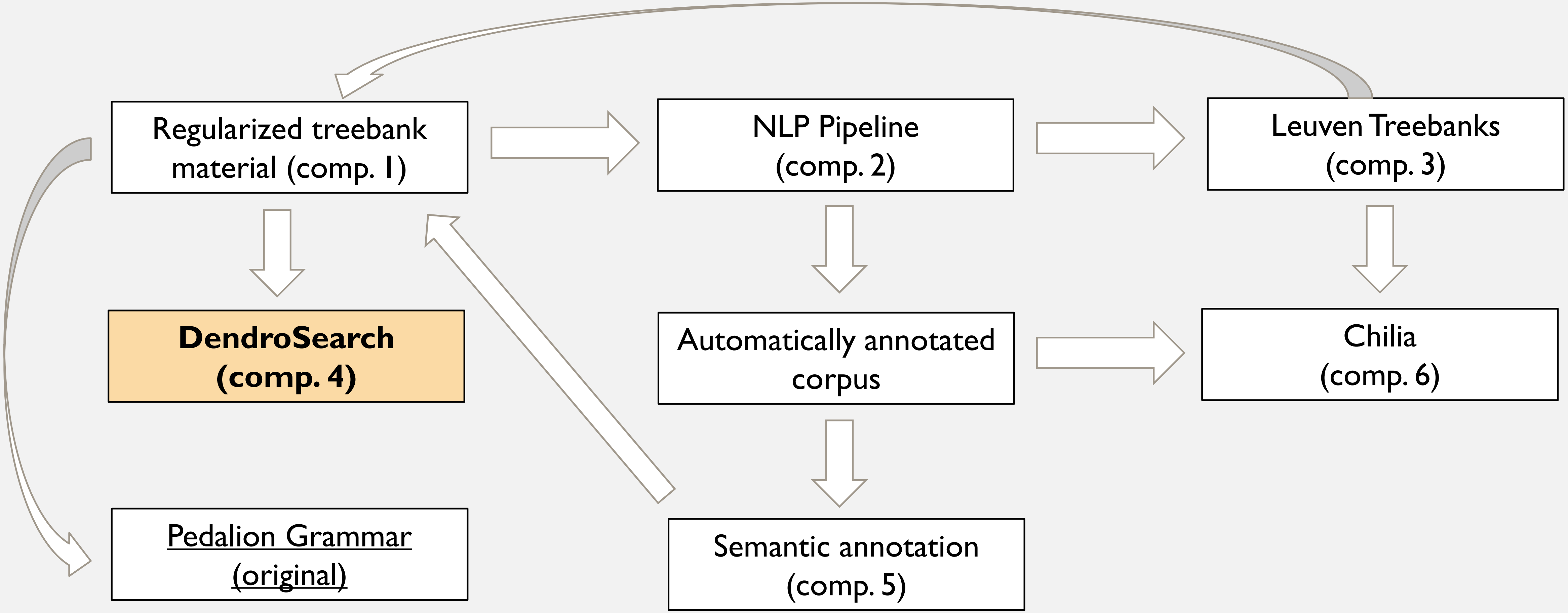
DELIVERABLE 3: LEUVEN TREEBANKS (2)



DELIVERABLE 3: LEUVEN TREEBANKS (3)



THE PEDALION ECOSYSTEM



Regularized treebank material (comp. 1)

NLP Pipeline (comp. 2)

Leuven Treebanks (comp. 3)

DendroSearch (comp. 4)

Automatically annotated corpus

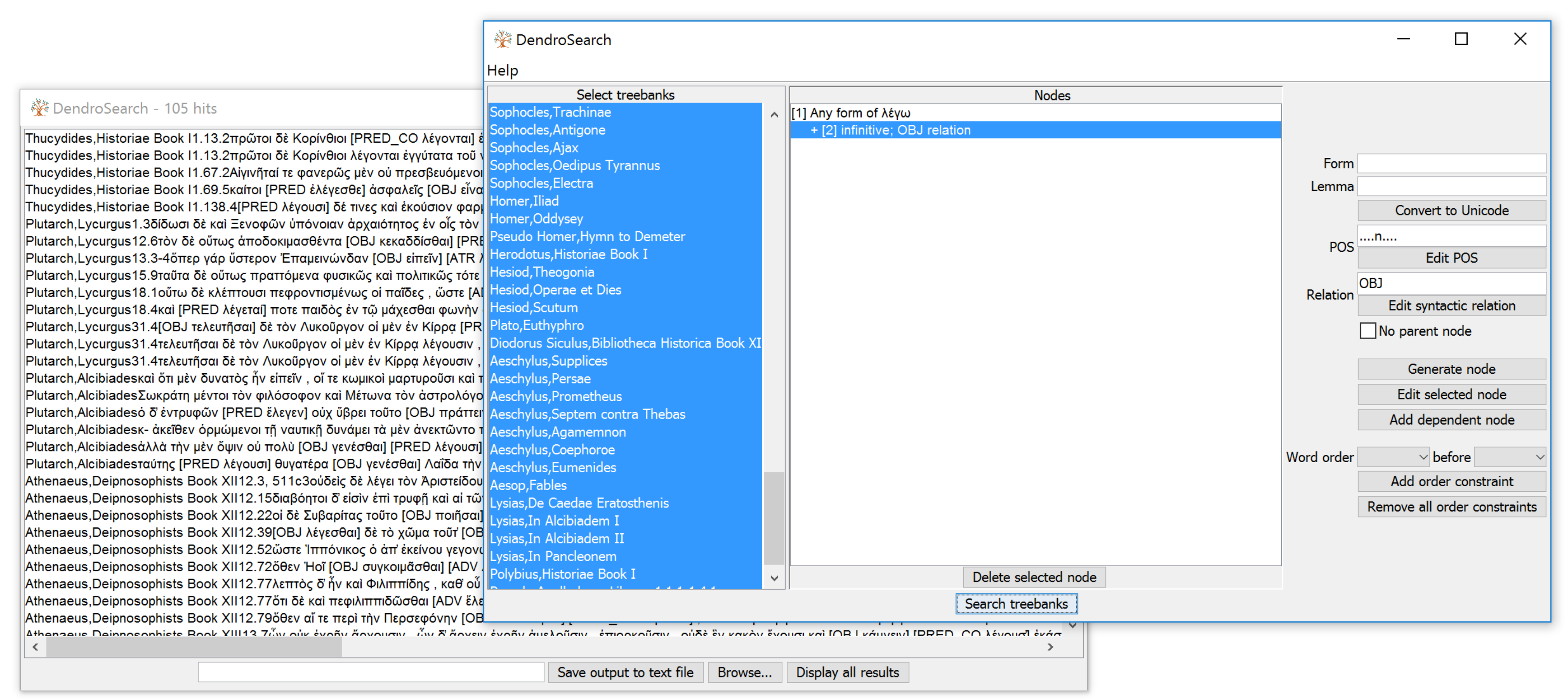
Chilia (comp. 6)

Pedalion Grammar (original)

Semantic annotation (comp. 5)

DELIVERABLE 4: DENDROSEARCH

- User-friendly tool to query the treebank material



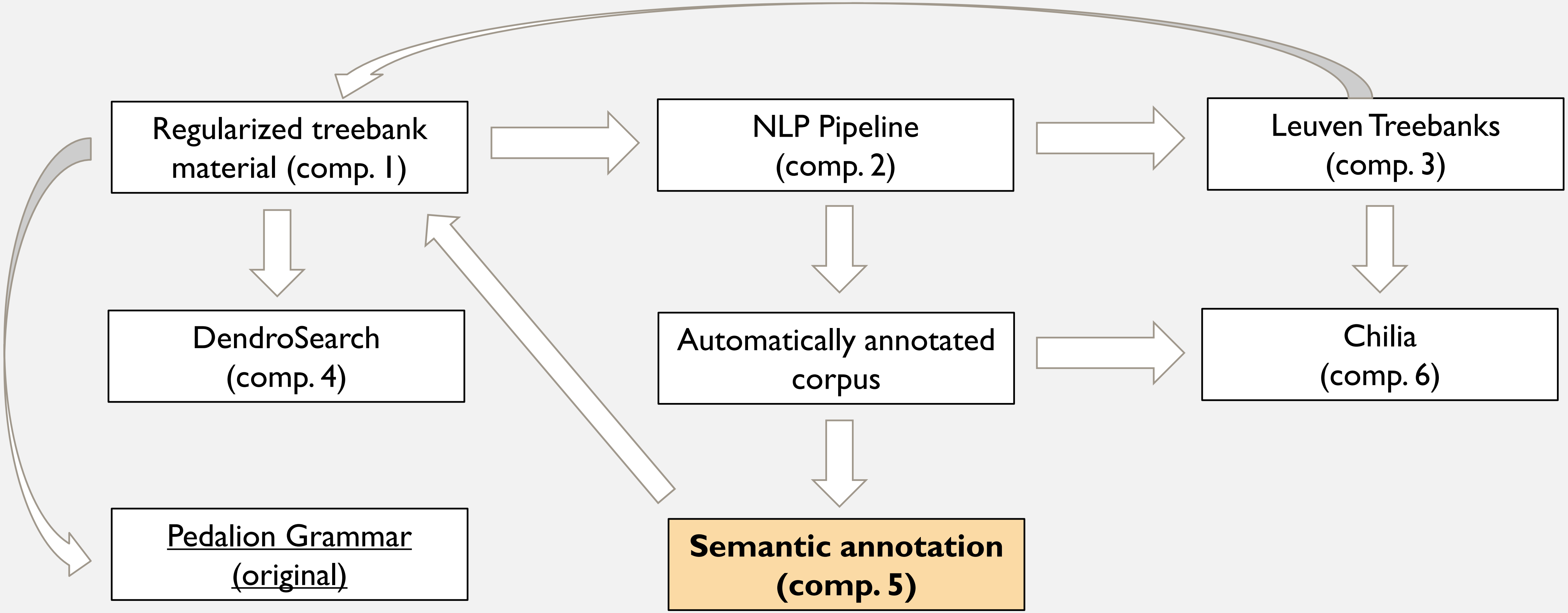
Thucydides,Historiae Book I1.13.2πρῶτοι δὲ Κορίνθιοι [PRED_CO λέγονται] ἐ
Thucydides,Historiae Book I1.13.2πρῶτοι δὲ Κορίνθιοι λέγονται ἐγγύτατα τοῦ ν
Thucydides,Historiae Book I1.67.2Αἰγινῆταί τε φανερῶς μὲν οὐ πρεσβευόμενοι
Thucydides,Historiae Book I1.69.5καίτοι [PRED ἐλέγεσθε] ἀσφαλεῖς [OBJ εἶνα
Thucydides,Historiae Book I1.138.4[PRED λέγουσι] δέ τινες καὶ ἑκούσιον φαρ
Plutarch,Lycurgus1.3δίδωσι δὲ καὶ Ξενοφῶν ὑπόνοιαν ἀρχαιότητος ἐν οἷς τὸν
Plutarch,Lycurgus12.6τὸν δὲ οὕτως ἀποδοκιμασθέντα [OBJ κεκαδδίσθαι] [PRE
Plutarch,Lycurgus13.3-4ὅπερ γάρ ὕστερον Ἐπαμεινώνδαν [OBJ εἰπεῖν] [ATR λ
Plutarch,Lycurgus15.9ταῦτα δὲ οὕτως πραττόμενα φυσικῶς καὶ πολιτικῶς τότε
Plutarch,Lycurgus18.1οὕτω δὲ κλέπτουσι πεφροντισμένως οἱ παῖδες , ὥστε [AI
Plutarch,Lycurgus18.4καὶ [PRED λέγεταί] ποτε παιδὸς ἐν τῷ μάχεσθαι φωνὴν
Plutarch,Lycurgus31.4[OBJ τελευτῆσαι] δὲ τὸν Λυκοῦργον οἱ μὲν ἐν Κίρρᾳ [PR
Plutarch,Lycurgus31.4τελευτῆσαι δὲ τὸν Λυκοῦργον οἱ μὲν ἐν Κίρρᾳ λέγουσιν ,
Plutarch,Lycurgus31.4τελευτῆσαι δὲ τὸν Λυκοῦργον οἱ μὲν ἐν Κίρρᾳ λέγουσιν ,
Plutarch,Alcibiadesκαὶ ὅτι μὲν δυνατὸς ἦν εἰπεῖν , οἵ τε κωμικοὶ μαρτυροῦσι καὶ τ
Plutarch,AlcibiadesΣωκράτη μέντοι τὸν φιλόσοφον καὶ Μέτωνα τὸν ἀστρολόγο
Plutarch,Alcibiadesὁ δ᾽ ἐντρυφῶν [PRED ἔλεγεν] οὐχ ὕβρει τοῦτο [OBJ πράττει
Plutarch,Alcibiadesκ- ἀκεῖθεν ὁρμώμενοι τῇ ναυτικῇ δυνάμει τὰ μὲν ἀνεκτῶντο τ
Plutarch,Alcibiadesἀλλὰ τὴν μὲν ὄψιν οὐ πολὺ [OBJ γενέσθαι] [PRED λέγουσι]
Plutarch,Alcibiadesταύτης [PRED λέγουσι] θυγατέρα [OBJ γενέσθαι] Λαΐδα τὴν
Athenaeus,Deipnosophists Book XII12.3, 511c3οὐδεὶς δὲ λέγει τὸν Ἀριστείδου
Athenaeus,Deipnosophists Book XII12.15διαβόητοι δ᾽ εἰσὶν ἐπὶ τρυφῇ καὶ αἱ τῶ
Athenaeus,Deipnosophists Book XII12.22οἱ δὲ Συβαρίτας τοῦτο [OBJ ποιῆσαι]
Athenaeus,Deipnosophists Book XII12.39[OBJ λέγεσθαι] δὲ τὸ χῶμα τοῦτ᾽ [OB
Athenaeus,Deipnosophists Book XII12.52ὥστε Ἱππόνικος ὁ ἀπ᾽ ἐκείνου γεγονι
Athenaeus,Deipnosophists Book XII12.72ὅθεν Ἠοῖ [OBJ συγκοιμᾶσθαι] [ADV
Athenaeus,Deipnosophists Book XII12.77λεπτὸς δ᾽ ἦν καὶ Φιλιππίδης , καθ᾽ οὗ
Athenaeus,Deipnosophists Book XII12.77ὅτι δὲ καὶ πεφιλιππιδῶσθαι [ADV ἔλε
Athenaeus,Deipnosophists Book XII12.79ὅθεν αἵ τε περὶ τὴν Περσεφόνην [OB
Athenaeus,Deipnosophists Book XIII13.7ὧν οὐκ ἐχρῆν ἄρχουσιν , ὧν δ᾽ ἄρχειν ἐχρῆν ἀμελοῦσιν , ἐπιορκοῦσιν , οὐδὲ ἂν κακὸν ἔχουσι καὶ [OBJ κάμνειν] [PRED_CO λέγουσι] ἑκάσ

Sophocles,Trachinae
Sophocles,Antigone
Sophocles,Ajax
Sophocles,Oedipus Tyrannus
Sophocles,Electra
Homer,Iliad
Homer,Oddysey
Pseudo Homer,Hymn to Demeter
Herodotus,Historiae Book I
Hesiod,Theogonia
Hesiod,Operae et Dies
Hesiod,Scutum
Plato,Euthyphro
Diodorus Siculus,Bibliotheca Historica Book XI
Aeschylus,Supplices
Aeschylus,Persae
Aeschylus,Prometheus
Aeschylus,Septem contra Thebas
Aeschylus,Agamemnon
Aeschylus,Coephoroe
Aeschylus,Eumenides
Aesop,Fables
Lysias,De Caedae Eratosthenis
Lysias,In Alcibiadem I
Lysias,In Alcibiadem II
Lysias,In Pancleonem
Polybius,Historiae Book I

[1] Any form of λέγω
+ [2] infinitive; OBJ relation

Form:
Lemma:
POS: ....n....
Relation: OBJ
☐ No parent node
Word order: before

THE PEDALION ECOSYSTEM



Regularized treebank material (comp. 1)

NLP Pipeline (comp. 2)

Leuven Treebanks (comp. 3)

DendroSearch (comp. 4)

Automatically annotated corpus

Chilia (comp. 6)

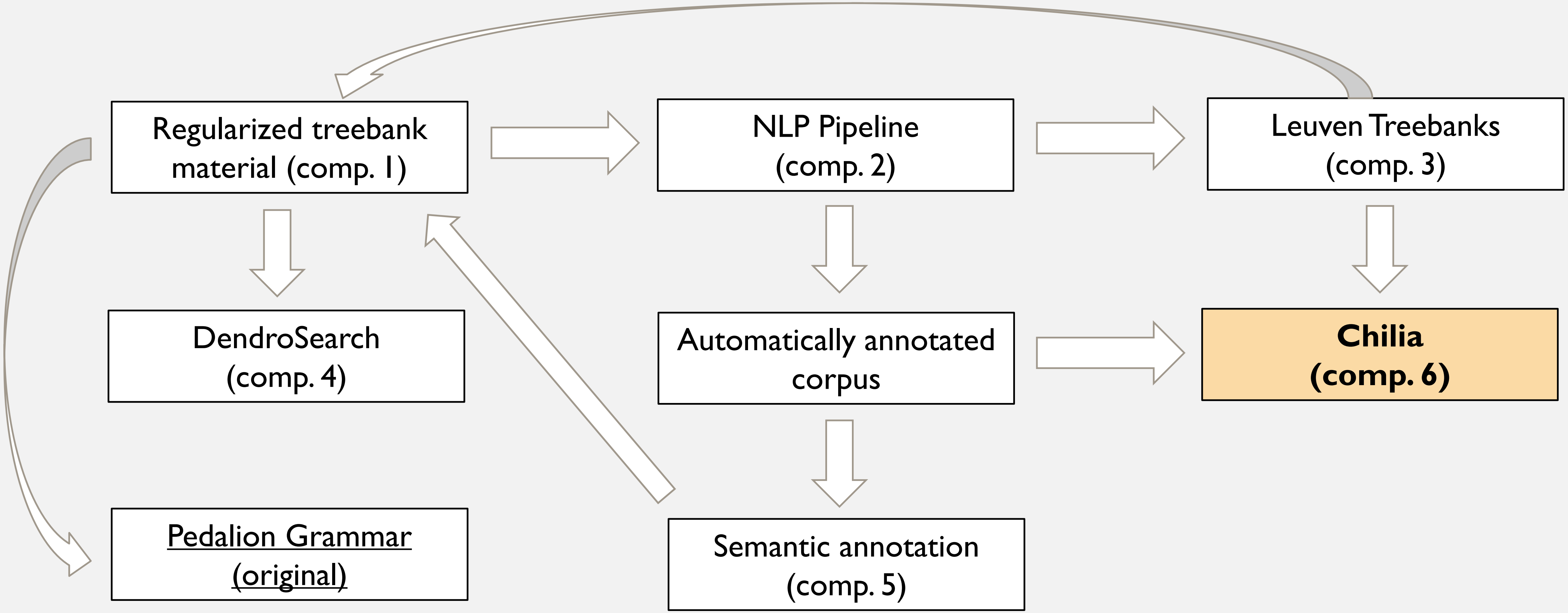
Pedalion Grammar (original)

Semantic annotation (comp. 5)

DELIVERABLE 5: SEMANTICS

- Semantic annotation on various levels
- On the word level:
  - Word vectors, using a large corpus (37 million tokens) as input material
    - E.g. extending queries DendroSearch with synonyms: X ἤμερας ~ ἔτη, ἐνιαυτούς etc.
  - Annotation of noun categories (e.g. animal, person, non-concrete etc.), verb categories (e.g. emotion, cognition, motion etc.), adjective categories (e.g. quantifier/qualifier), also using word vectors as input
- On the phrasal level: semantic roles (manually so far, but experimenting with automatic annotation), see Pedalion roles
- See Swaelens 2019

THE PEDALION ECOSYSTEM



Regularized treebank material (comp. 1)

NLP Pipeline (comp. 2)

Leuven Treebanks (comp. 3)

DendroSearch (comp. 4)

Automatically annotated corpus

Chilia (comp. 6)

Pedalion Grammar (original)

Semantic annotation (comp. 5)

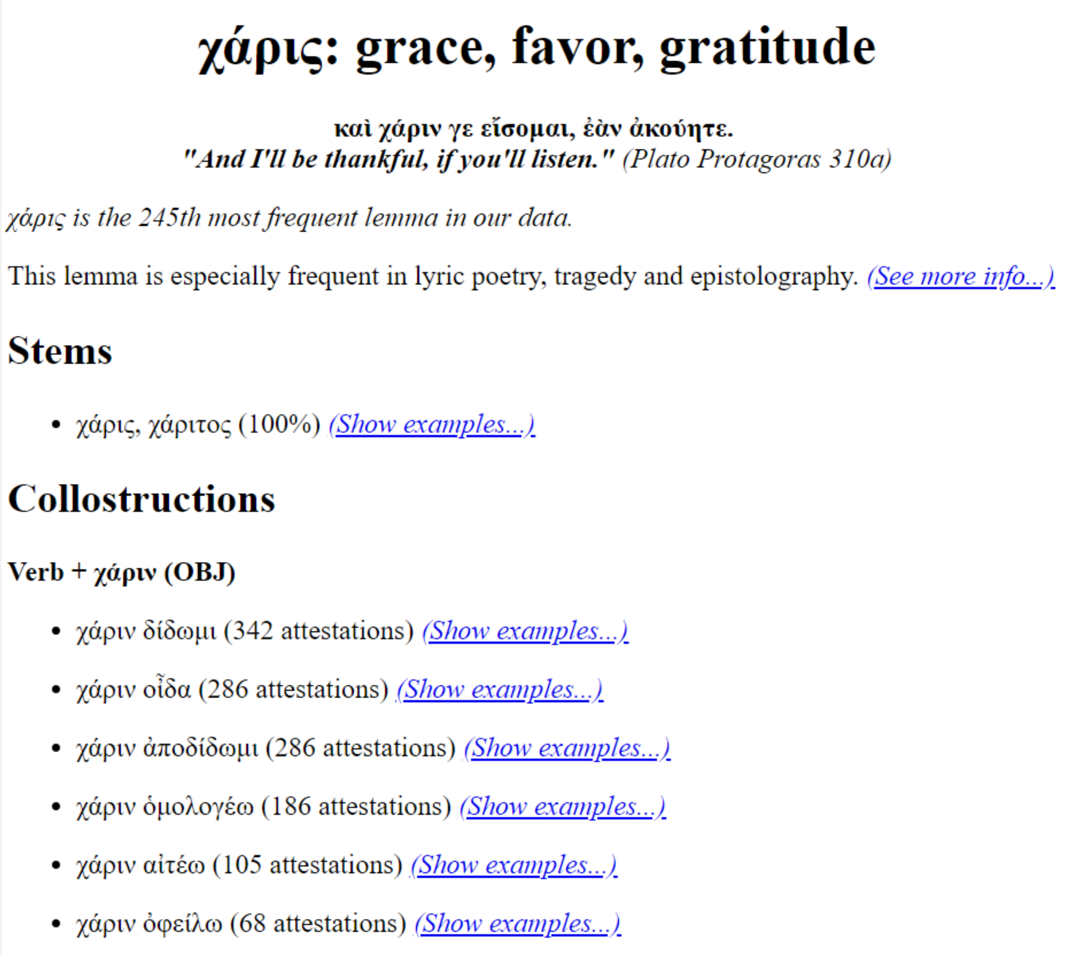
DELIVERABLE 6: CHILIA (1)

- A list of 1000 ‘key words’ of Ancient Greek, corpus-based
- Context: pedagogical material for pupils in high schools and university students without previous knowledge of Greek
- Every word is illustrated with an authentic sentence containing only words from this Chilia list:
  - “ἦν γάρ **ποτε** χρόνος, ὦ Ἀθηναῖοι, ὅτε τείχη καὶ ναῦς οὐκ ἐκεκτήμεθα” [**“Once** there was a time, Athenians, when we had neither walls nor a fleet”].

DELIVERABLE 6: CHILIA (2)

- For an online version, we envision a more dynamic word list
- Filtering by genre, e.g. key words of medical texts: κοιλία (belly), πυρετός (fever), ὀδύνη (pain), σιτίον (food), τρίβω (rub) etc. – of historical texts: στρατιά (army), στράτευμα (army), ἱππεύς (horseman), στρατεύω (wage war), ὁπλίτης (hoplite) etc.
- 'Identikit' of each word: e.g. most typical genres, diachronic course, collostructions
- Links to *Logeion*

DELIVERABLE 6: CHILIA (3)



χάρις: grace, favor, gratitude

καὶ χάριν γε εἴσομαι, ἐὰν ἀκούητε.
"And I'll be thankful, if you'll listen." *(Plato Protagoras 310a)*

χάρις is the 245th most frequent lemma in our data.

This lemma is especially frequent in lyric poetry, tragedy and epistolography. *(See more info...)*

Stems

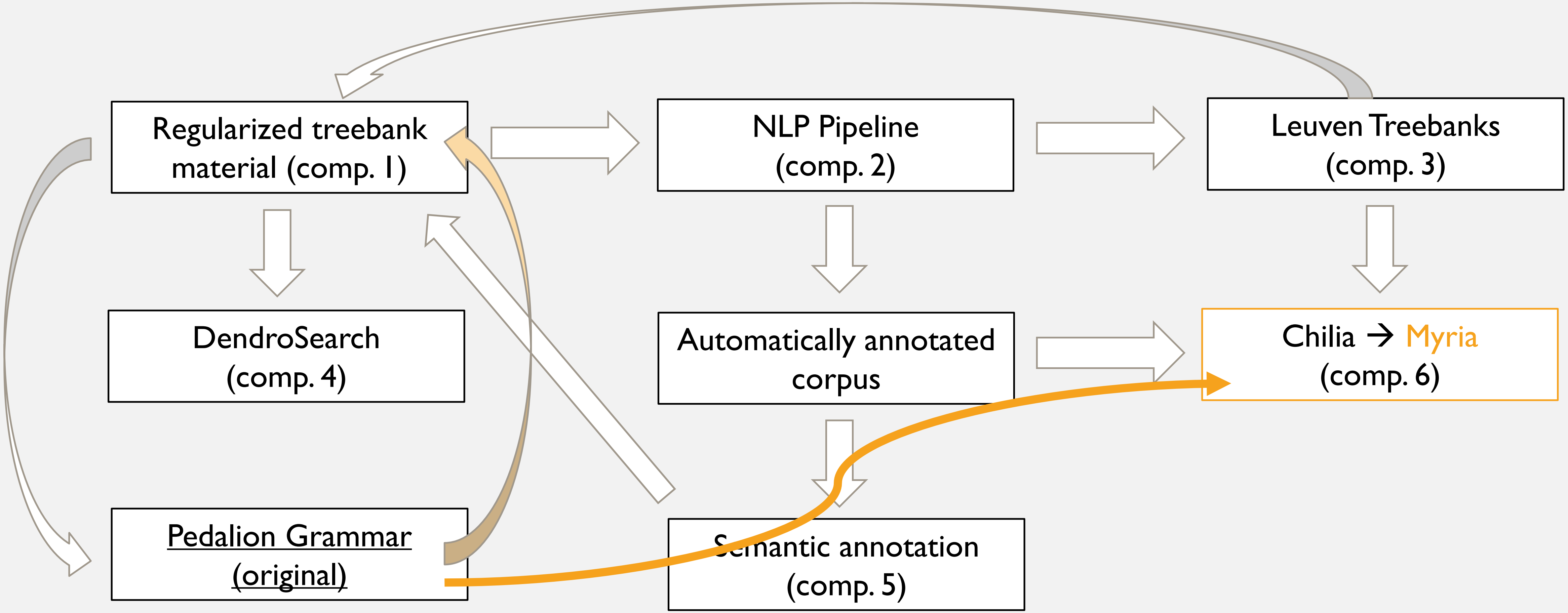
- χάρις, χάριτος (100%) *(Show examples...)*

Collostructions

Verb + χάριν (OBJ)

- χάριν δίδωμι (342 attestations) *(Show examples...)*
- χάριν οἶδα (286 attestations) *(Show examples...)*
- χάριν ἀποδίδωμι (286 attestations) *(Show examples...)*
- χάριν ὁμολογέω (186 attestations) *(Show examples...)*
- χάριν αἰτέω (105 attestations) *(Show examples...)*
- χάριν ὀφείλω (68 attestations) *(Show examples...)*

PEDALION: FURTHER POSSIBILITIES



Regularized treebank material (comp. 1)

NLP Pipeline (comp. 2)

Leuven Treebanks (comp. 3)

DendroSearch (comp. 4)

Automatically annotated corpus

Chilia → Myria (comp. 6)

Pedalion Grammar (original)

Semantic annotation (comp. 5)