
UNIVERSITY OF OXFORD
FACULTY OF CLASSICS
LEVERHULME TRUST
Interfacing: The PARSEME Ancient Greek Corpus
A hands-on workshop
Churchill College, Cambridge
11 March 2024

Programme

9:00–10:30am Session 1 (Annotation guidelines for a corpus language)

10:30–11:00 Break

11:00–12:30 Session 2 (Literary classical Attic and inter-annotator scores)

12:30–2:00pm Sandwich lunch (hall)

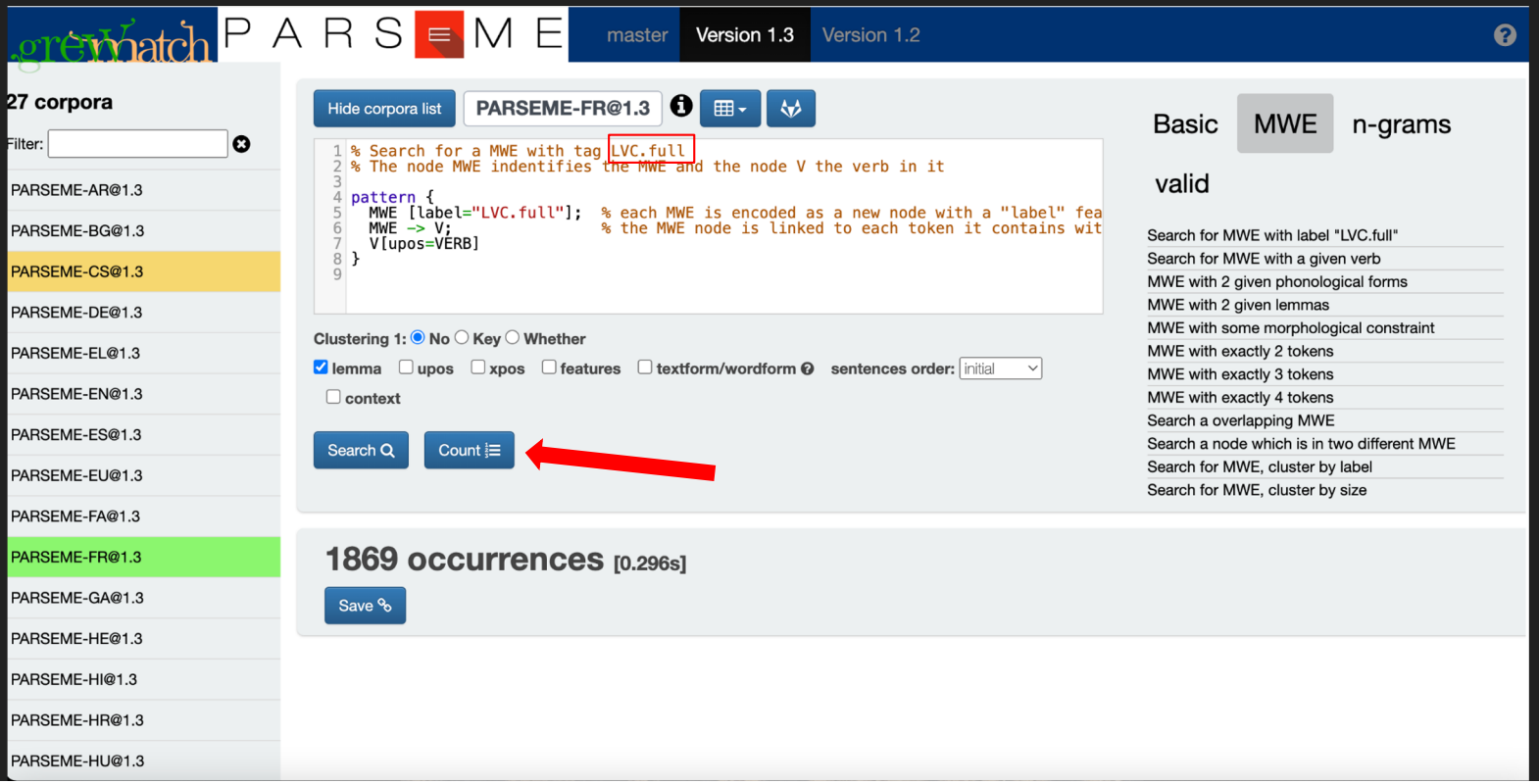
2:00–3:30pm Session 3 (Polybios and the papyri in UD pipe)

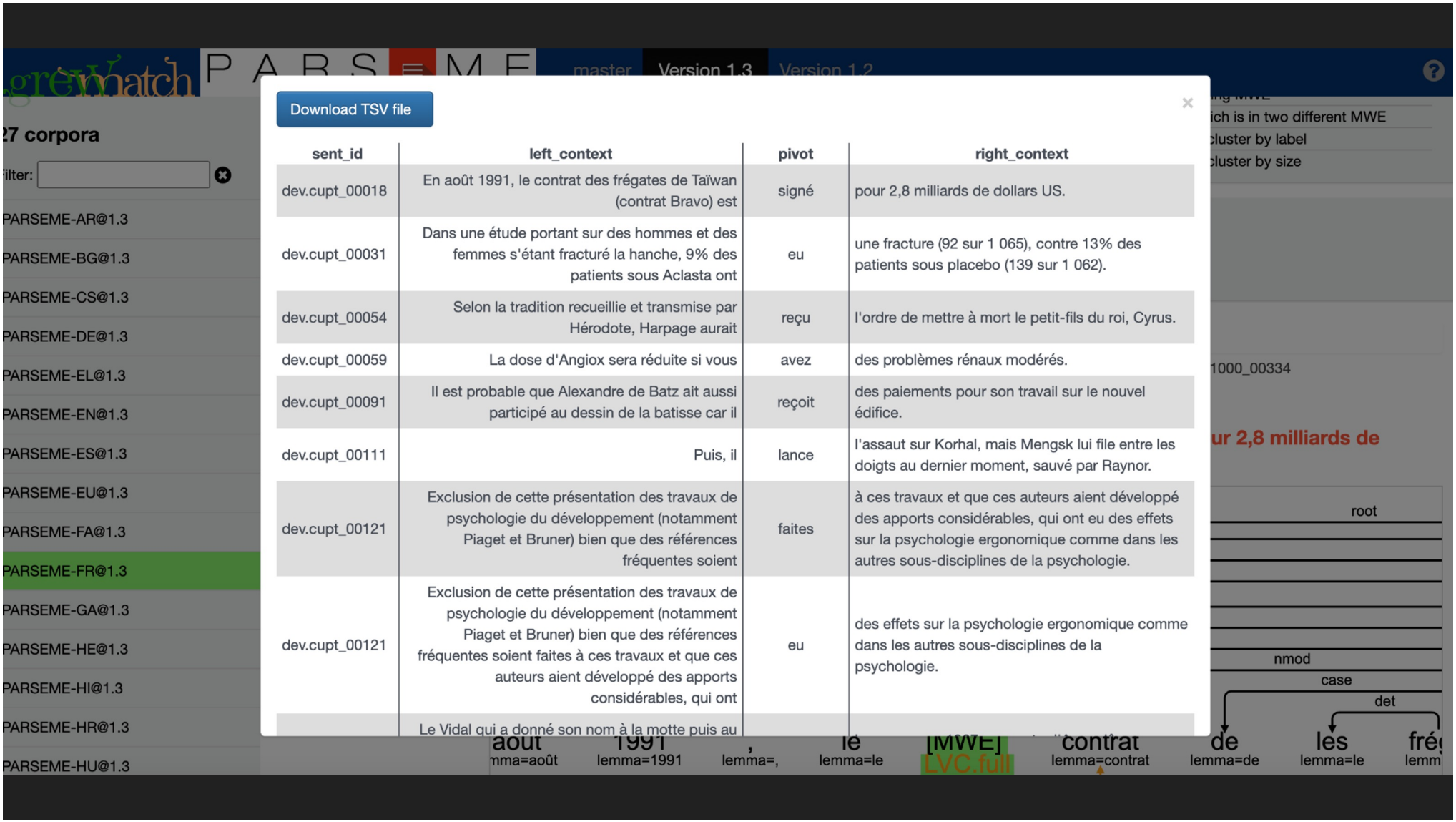
3:30–4:00pm Break

4:00–5:30pm Session 4 (Wikipedia Light-verb constructions & nominal multi-word constructions)

What is PARSEME?

https://parseme.grew.fr/?custom=65e0996707808# GrewMatch





Download TSV file

sent_id	left_context	pivot	right_context
dev.cupt_00018	En août 1991, le contrat des frégates de Taïwan (contrat Bravo) est	signé	pour 2,8 milliards de dollars US.
dev.cupt_00031	Dans une étude portant sur des hommes et des femmes s'étant fracturé la hanche, 9% des patients sous Aclasta ont	eu	une fracture (92 sur 1 065), contre 13% des patients sous placebo (139 sur 1 062).
dev.cupt_00054	Selon la tradition recueillie et transmise par Hérodote, Harpage aurait	reçu	l'ordre de mettre à mort le petit-fils du roi, Cyrus.
dev.cupt_00059	La dose d'Angiox sera réduite si vous	avez	des problèmes rénaux modérés.
dev.cupt_00091	Il est probable que Alexandre de Batz ait aussi participé au dessin de la batisse car il	reçoit	des paiements pour son travail sur le nouvel édifice.
dev.cupt_00111	Puis, il	lance	l'assaut sur Korhal, mais Mengsk lui file entre les doigts au dernier moment, sauvé par Raynor.
dev.cupt_00121	Exclusion de cette présentation des travaux de psychologie du développement (notamment Piaget et Bruner) bien que des références fréquentes soient	faites	à ces travaux et que ces auteurs aient développé des apports considérables, qui ont eu des effets sur la psychologie ergonomique comme dans les autres sous-disciplines de la psychologie.
dev.cupt_00121	Exclusion de cette présentation des travaux de psychologie du développement (notamment Piaget et Bruner) bien que des références fréquentes soient faites à ces travaux et que ces auteurs aient développé des apports considérables, qui ont	eu	des effets sur la psychologie ergonomique comme dans les autres sous-disciplines de la psychologie.
	Le Vidal qui a donné son nom à la motte puis au		

Why an Ancient Greek corpus?

- The issue of dictionaries and authoritative dictionaries
- Literae Humaniores Mods
  - parse / describe / comment (form, structure, meaning)
- Message string on Classics Liverpool list …

(LYSIAS 3.15-18)

Answer the following questions:

a) Parse, then describe and comment on the use of the following participles: συνδραμόντων (6), λεγομένων (8), ἐρωτηθέντες (14).

b) Parse, then describe and comment on the use of the following perfect form: μεμαρτύρηται (3).

c) Parse and describe, and comment on, the use of tense and aspect in the following forms: ἦγον (5), περιιδεῖν (11).

d) Parse, describe, and comment on, the use of the following cases: βίᾳ (5), αὐτοῖς (10), μάχης (15), κεφαλὰς (21).

e) Lysias' style has been described as follows: 'The dominant impression created is one of artlessness' (C. Carey). Do you agree? Support your argument with two examples taken from the passage above.

Request for High-Quality Literal Translations of Latin Texts for Large Language Model Research

PR

Classicists <CLASSICISTS@liverpool.ac.uk> on behalf of Paul Rosu <Paul

To: CLASSICISTS@liverpool.ac.uk

Sat 03/02/2024 12:12

Dear Members of the Classics Community,

I am an undergraduate at Duke University, currently working to fine-tune a large language model to produce free, high-quality, literal English translations of Latin. My goal is to help researchers from fields outside of Classics to access Latin texts that might otherwise be inaccessible to them.

The success of the project depends on the accuracy of the translations used to train the model. I would be most grateful for your assistance in developing this set of training data.

I am seeking volunteers with excellent Latin who are willing to offer English translations of one or more passages of Latin (along with the Latin passage). Any author or text will

Why an Ancient Greek corpus?

Let’s square a circle!

Haspelmath 2010, p. 665 ‘**Comparative concepts** are concepts created by comparative linguists for the specific purpose of crosslinguistic comparison. Unlike descriptive categories, they are not part of particular language systems and are not needed by descriptive linguists or by speakers. They are not psychologically real, and they cannot be right or wrong. They can only be more or less well suited to the task of permitting crosslinguistic comparison. They are often labeled in the same way as descriptive categories, but they stand in a many-to-many relationship with them (§ 9). Comparative concepts are universally applicable, and they are defined on the basis of other universally applicable concepts: universal conceptual-semantic concepts, general formal concepts, and other comparative concepts. Comparative concepts have often been used implicitly in the typological literature, but there has not been any detailed and explicit discussion of the difference between comparative concepts and language-particular descriptive categories.’

Corpus preparation

- Licence → select version of the text

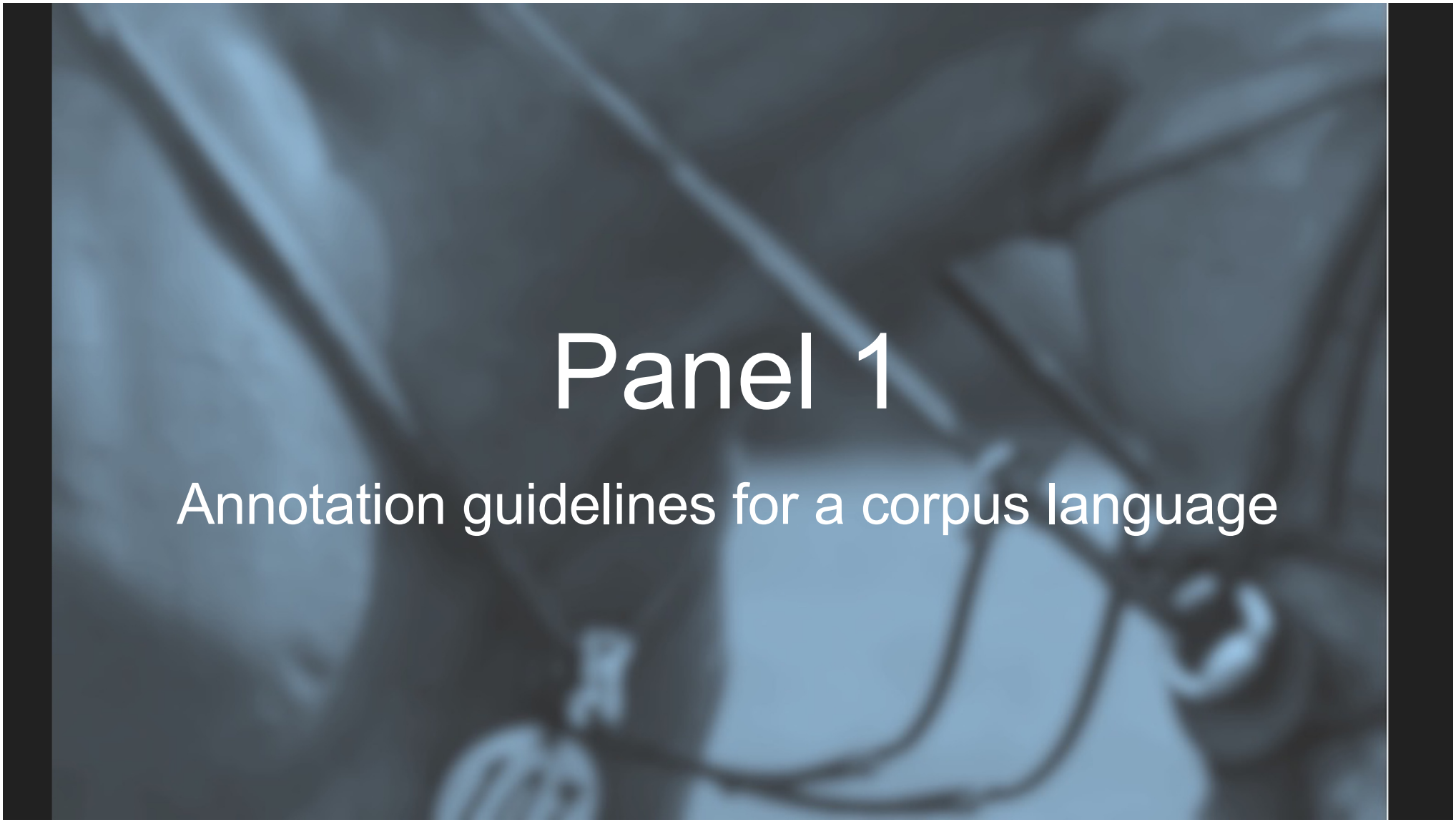
Creative Commons Licence

- Text type → parliamentary speeches, newspaper articles, interviews currently in other corpora

What is literature?

- UD model → Perseus

Limitations of the model (e.g. synchronic variability and diachronic change)


Panel 1
Annotation guidelines for a corpus language

Annotation guidelines (i)

https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/index.php?page=home

- Words and tokens → NLP approach, thus consider UniDive model selected → example of apostrophes fused to words in deletion contexts
- Verbal multi-word expressions → new initiative currently, see Panel 4
- Problem 1: Syntactic variants
- Problem 2: Collocations and the notion of lexicalisation
- Problem 3: Metaphors and idiomaticity
- Problem 4: TODO category (transitive verb + adverb; internal and cognate objects; lexical passives & Co.)
- Problem 5: The concept of decision trees

Words and tokens – UD pipe

τ᾽ἦν τ᾽ἦ

```
# text = τ᾽ἦν
1	τ	ε	PRON	g--------	_	2	advmod	_	SpaceAfter=No|TokenRange=0:1
2	᾽ἦν	εἰμί	VERB	v3siia---	Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act	0	root	_	SpaceAfter=No|TokenRange=1:4
```

```
# text = τ᾽ ἦν
1	τ᾽	τ᾽	ADV	d--------	_	2	advmod	_	TokenRange=0:2
2	ἦν	εἰμί	AUX	v3siia---	Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act	0	root	_	SpaceAfter=No|TokenRange=3:5
```

τ᾽ἦσθα τ᾽ ἦσθα

```
# text = τ᾽ἦσθα
1	τ	ε	PRON	g--------	_	2	advmod	_	SpaceAfter=No|TokenRange=0:1
2	᾽ἦσθα	᾽ἦμι	VERB	v2siia---	Aspect=Imp|Mood=Ind|Number=Sing|Person=2|Tense=Past|VerbForm=Fin|Voice=Act	0	root	_	SpaceAfter=No|TokenRange=1:6
```

```
# text = τ᾽ ἦσθα
1	τ᾽	τ᾽	ADV	d--------	_	2	advmod	_	TokenRange=0:2
2	ἦσθα	εἰμί	AUX	v2siia---	Aspect=Imp|Mood=Ind|Number=Sing|Person=2|Tense=Past|VerbForm=Fin|Voice=Act	0	root	_	SpaceAfter=No|TokenRange=3:7
```

What about crasis phenomena? E.g. τἀληθῆ τἄλλα & τοὔπισθεν

Problem 1: Syntactic variants

https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/index.php?page=010_Definitions_and_scope/030_Syntactic_variants_of_VMWEs

Fleischman 2000, p. 34 The term ‘text language’ is intended to reflect the fact that the linguistic activity of such languages is amenable to scrutiny only insofar as it has been constituted in the form of extant *texts*, which we might think of as its ‘native speakers’, even if we can’t interrogate them in quite the same way as we can native speakers of living languages.

Problem 2: Collocations and the notion of lexicalisation

Lexicalisation → to take someone by surprise

Some components of such compulsory arguments may be **lexicalized**, that is, always realized by the same lexemes. Here, *by surprise* is lexicalized while *someone* is not. The head verb of a VMWE is always considered lexicalized. When it can be replaced by another verb, like in *to* ***make****/take a decision*, we consider that these are two different VMWEs, although possibly synonymous.

Vs Boye 2023, p. 274 Lexical elements (meanings, morphemes, words and constructions) are by convention potentially discursively primary: they can, but need not, be the attentional main point of a syntagm.

Collocation

We understand **collocations** as combinations of words whose idiosyncrasy is **purely statistical**. In other words, tokens in collocations tend to co-occur with each other more often than expected by chance, but they show no substantial orthographic, morphological, syntactic and (most notably) semantic idiosyncrasy. In this way we **oppose** MWEs to collocations.

Baayen 2009, pp. 904–907 type count and realized productivity, rate of expansion (and hapaces) and expanding productivity, ratio of hapaces to total of tokens in a category and potential productivity

Problem 3: Metaphors and idiomaticity

- To take the bull by the horns
- To set the world on fire
- To put all one's eggs in one basket

Lakoff and Johnson 1988, p. 3 metaphor is pervasive in everyday life, not just in language but in thought and action (conceptual metaphor, e.g. argument is war)

Charteris-Black 2021, p. 6 Metaphors contribute to the moral framing of a situation in such a way that we become biased towards one form of action over another, and they provide insight into the moral framing of our actions.

Lausberg and Orton 1998, pp. 250ff. the *metaphora* (Quint. *Inst.* 8.6.8), ~E'ta<f>opcX (Tryph. *Trop.* iii p.191,24), *trans/atio* (Quint. *Inst.* 8.6.4; *Rhet.Her.* 4.45) is explained as the *brevitas-form* of the comparison

Tindale 2010, p. 31a range of metaphoric devices such as analogies and similes

PARSEME 1.3 These expressions, however, were probably constructed for the needs of one article/poem only and are not sufficiently established in the common vocabulary to be considered VMWEs.

Problem 4: TODO category

Examples taken from Lysias, Speech 1:

- transitive verb + adverb
  - ἔχω + adverb
  - τίθημι / κεῖμαι + adverb
- internal and cognate objects
  - ἁρμάτημα ἐξαμαρτάνειν
- lexical passives & Co. (cf. Gross 1998)
  - συγγνώμης τυγχάνω
- Syntactic nominalisations?!
  - κακοῖς συνέχομαι
- S + γίγνομαι (cf. Modern Greek and Jímenez López 2021)
  - σπονδαὶ γίγνονται

Problem 5: Decision trees

Generic decision tree:

- As long as the tests are based on structure, we can usually apply them without issues
- BUT what about statistical relevance?
- BUT what about judging something ‘ungrammatical / unidiomatic’?

↳Apply test S.1 - [**1HEAD**: Unique verb as functional syntactic head of the whole?]
- ↳ **NO ⇒** Apply the VID-specific tests ⇒ *VID tests positive?*
  - ↳ **YES ⇒** Annotate as a VMWE of category **VID**
  - ↳ **NO ⇒** It is not a VMWE, **exit**
- ↳ **YES ⇒** Apply test S.2 - [**1DEP**: *Verb v has exactly one lexicalized dependent d?*]
  - ↳ **NO ⇒** Apply the VID-specific tests ⇒ *VID tests positive?*
    - ↳ **YES ⇒** Annotate as a VMWE of category **VID**
    - ↳ **NO ⇒** It is not a VMWE, **exit**
  - ↳ **YES ⇒** Apply test S.3 - [**LEX-SUBJ**: *Lexicalized subject?*]
    - ↳ **YES ⇒** Apply the VID-specific tests ⇒ *VID tests positive?*
      - ↳ **YES ⇒** Annotate as a VMWE of category **VID**
      - ↳ **NO ⇒** It is not a VMWE, **exit**
    - ↳ **NO ⇒** Apply test S.4 - [**CATEG**: *What is the morphosyntactic category of d?*]
      - ↳**Reflexive clitic** ⇒ Apply IRV-specific tests ⇒ *IRV tests positive?*
        - ↳ **YES ⇒** Annotate as a VMWE of category **IRV**
        - ↳ **NO ⇒** It is not a VMWE, **exit**
      - ↳**Particle** ⇒ Apply VPC-specific tests ⇒ *VPC tests positive?*
        - ↳ **YES ⇒** Annotate as a VMWE of category **VPC.full** or **VPC.semi**
        - ↳ **NO ⇒** It is not a VMWE, **exit**
      - ↳**Verb with no lexicalized dependent** ⇒ Apply MVC-specific tests ⇒ *MVC tests positive?*
        - ↳ **YES ⇒** Annotate as a VMWE of category **MVC**
        - ↳ **NO ⇒** Apply the VID-specific tests ⇒ *VID tests positive?*
          - ↳ **YES ⇒** Annotate as a VMWE of category **ID**
          - ↳ **NO ⇒** It is not a VMWE, **exit**
      - ↳**Extended NP** ⇒ Apply LVC-specific decision tree ⇒ *LVC tests positive?*
        - ↳ **YES ⇒** Annotate as a VMWE of category **LVC**
        - ↳ **NO ⇒** Apply the VID-specific tests ⇒ *VID tests positive?*
          - ↳ **YES ⇒** Annotate as a VMWE of category **VID**
          - ↳ **NO ⇒** It is not a VMWE, **exit**
      - ↳**Another category** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
        - ↳ **YES ⇒** Annotate as a VMWE of category **VID**
        - ↳ **NO ⇒** It is not a VMWE, **exit**

Categories to be annotated

https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/index.php?page=030_Categories_of_VMWEs

- LVC.full
- LVC.cause
- VID
- MVC
- TODO (see Ancient Greek specific rules!)
- (IAV, VPC) (only post-classical)
- NOT VMWE

LVC

- Abstract
- Predicative

→ polysemy, instinct, …

- Light verb
- Reduction

→ redundancy?

LVC-specific decision tree:

↳Apply test LVC.0 - [**N-ABS**: *Is the noun abstract?*]
- ↳ **NO ⇒** It is not an LVC, exit
- ↳ **YES or UNSURE ⇒** Apply test LVC.1 - [**N-PRED**: *Is the noun predicative?*]
  - ↳ **NO ⇒** It is not an LVC, exit
  - ↳ **YES or UNSURE ⇒** Apply test LVC.2 - [**V-SUBJ-N-ARG**: *Is the subject of the verb a semantic argument of the noun?*]
    - ↳ **YES or UNSURE ⇒** Apply test LVC.3 - [**V-LIGHT**: *The verb only adds meaning expressed as morphological features?*]
      - ↳ **NO ⇒** It is not an LVC, exit
      - ↳ **YES ⇒** Apply test LVC.4 - [**V-REDUC**: *Can a verbless NP-reduction refer to the same event/state?*]
        - ↳ **NO ⇒** It is not an LVC, exit
        - ↳ **YES ⇒** It is an **LVC.full**
    - ↳ **NO ⇒** Apply test LVC.5 - [**V-SUBJ-N-CAUSE**: *Is the subject of the verb the cause of the noun?*]
      - ↳ **NO ⇒** It is not an LVC, exit
      - ↳ **YES ⇒** It is an **LVC.cause**

VID

περὶ πολλοῦ ποιέομαι

- τοῦ
- πολλῶν
- ἀγαθοῦ

What about examples such as:

- to take wing
- to take a picture (cf. Radimsky 2011)
- to spill the beans (cf. Mel'cuk 2023)

VID-specific decision tree:

Note: In this tree, a single YES to one of the tests is sufficient to decide that a candidate is a VID. Note however that this tree is to be applied only after it was referred to by the generic decision tree containing structural tests.

↳Apply test VID.1 - [**CRAN**: *Candidate contains cranberry word?*]
↳ **YES ⇒** It is a VID, exit.
↳ **NO ⇒** Apply test VID.2 - [**LEX**: *Regular replacement of a component ⇒ unexpected meaning shift?*]
↳ **YES ⇒** It is a VID, exit.
↳ **NO ⇒** Apply test VID.3 - [**MORPH**: *Regular morphological change ⇒ unexpected meaning shift?*]
↳ **YES ⇒** It is a VID, exit.
↳ **NO ⇒** Apply test VID.4 - [**MORPHSYNT**: *Regular morphosyntactic change ⇒ unexpected meaning shift?*]
↳ **YES ⇒** It is a VID, exit.
↳ **NO ⇒** Apply test VID.5 - [**SYNT**: *Regular syntactic change ⇒ unexpected meaning shift?*]
↳ **YES ⇒** It is a VID, exit.
↳ **NO ⇒** It is not a VID, exit

Gaps in the GRC guidelines

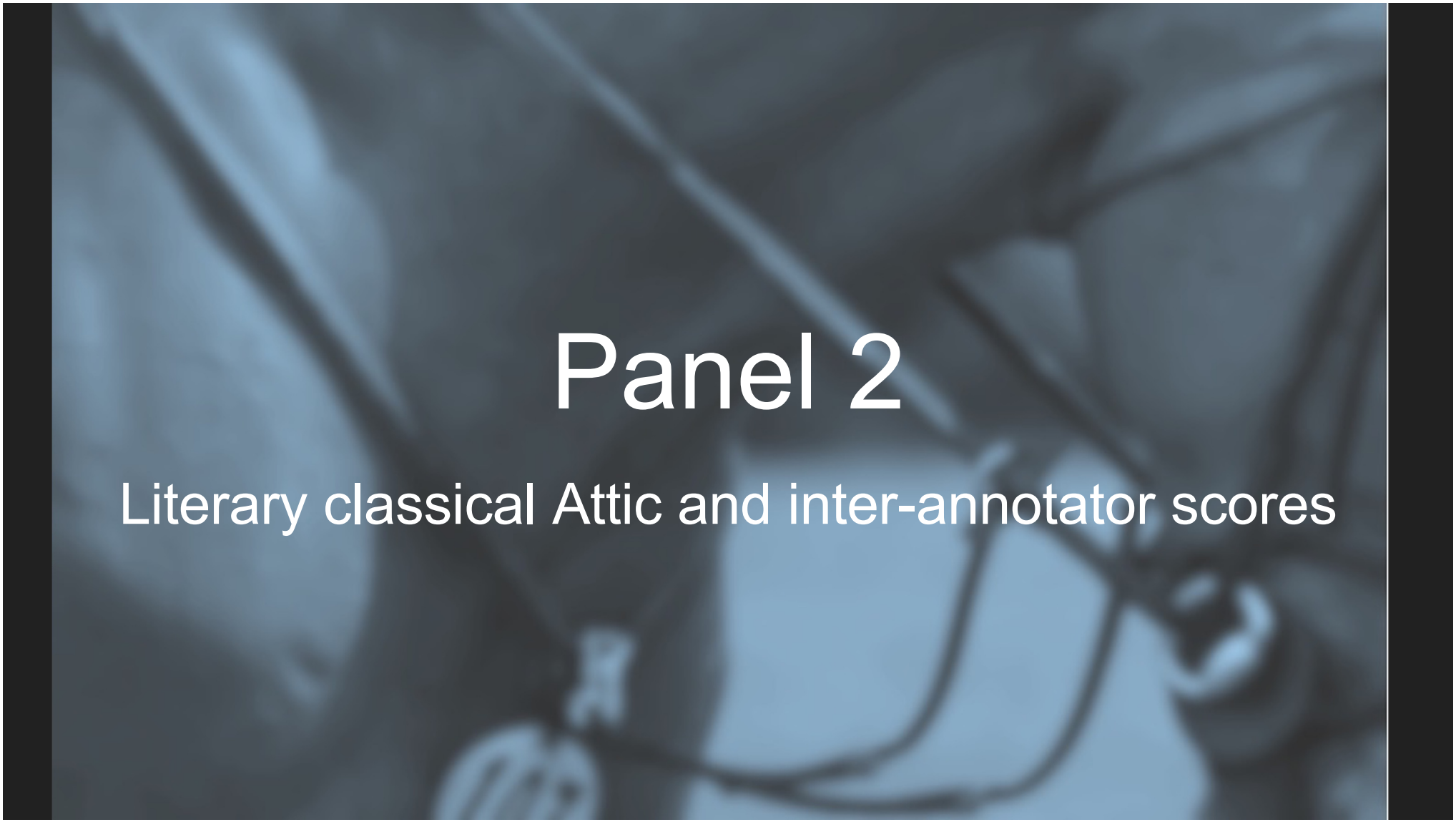
LINK for collaborative document: https://docs.google.com/document/d/1XfKVYTpld_ulDhQ-ECd0YWdhXye1PhywkJoxFx6r_Ik/edit?usp=sharing (also in booklet and via email)

Some ideas are in the guidelines already:

- 1.4 Fully saturated phrase
- 1.4 MWEs containing verbs but functioning as adverbials, adjectives or nominals that are not meaning-preserving variants
- (5.2) Test LVC.4 - [V-REDUC] - Verb reduction]
- 5.2 Therefore, you should NOT annotate as LVC.cause constructions involving …
- 5.2 Selection of the verb (i.e. reverse selection by the noun)
- 5.5 & 5.7 → see below (examples missing due to post-classical development!)

Examples entirely missing:

- 1.6 [collocation] Some combinations happen to be very frequent and are perceived as "frozen" (cf. ἀκριβῶς εἰδέναι, Bentein 2019, p. 147)
- 1.7 Metaphors
- 5.1 Test S.2 - [1DEP] - Single dependent → NO option
- 5.1 Test S.3 - [LEX-SUBJ] - Lexicalized subject (* cf. Homer Iliad 18.247 πάντας γὰρ ἔχε τρόμος)
- 5.1 Test S.4 - [CATEG] - Category of the dependent → extended NP
- 5.3 examples of types of VIDs
- **5.3**Test VID.1 - [CRAN] - Cranberry word → NO option
- **5.3** Test VID.2 - [LEX] - Lexical inflexibility → perhaps just different word instead of ‘letter’
- 5.6 Test MVC.13 - [V-LEX] Lexical inflexibility


Panel 2
Literary classical Attic and inter-annotator scores

Annotation practice

Text 1: Lysias, Speech 1 (On the murder of Eratosthenes)

Text 2: Xenophon, Anabasis, Book 1

Text 3: Plato, Republic, Book 1

Goal: Each text should be annotated for VMWE by at least 3 people in order to allow for inter-annotator comparison.

Document for results: https://docs.google.com/document/d/1XfKVYTpId_uIDhQ-ECd0YWdhXye1PhywkJoxFx6r_Ik/edit?usp=sharing (also in your booklet and in email)

Caveats

This is a collaboration with Natural Language Processing, thus

(1) [cf. LVC] we are forcing universality in places where it does not exist;

→ we are ignoring extra-linguistic indices (which would elucidate that there is in fact no redundancy in the Greek lexicon when it comes to MWEs) (e.g. Rusten 2020);

(1) [cf. LVC] we are using always the same meaning of the noun that we settle on and run through the tests;

→ there is the risk of circularity in a corpus language as we are relying on dictionaries that were in turn built based on these texts and by using contextual cues to disambiguate meaning; we are also to an extent ignoring polysemy and / or homonymy;

(1) [cf. LVC] we want to avoid gaps in the annotation as much as possible in order to train a machine on this eventually;

→ the issue of canonical forms not being attested and the like plays into this.

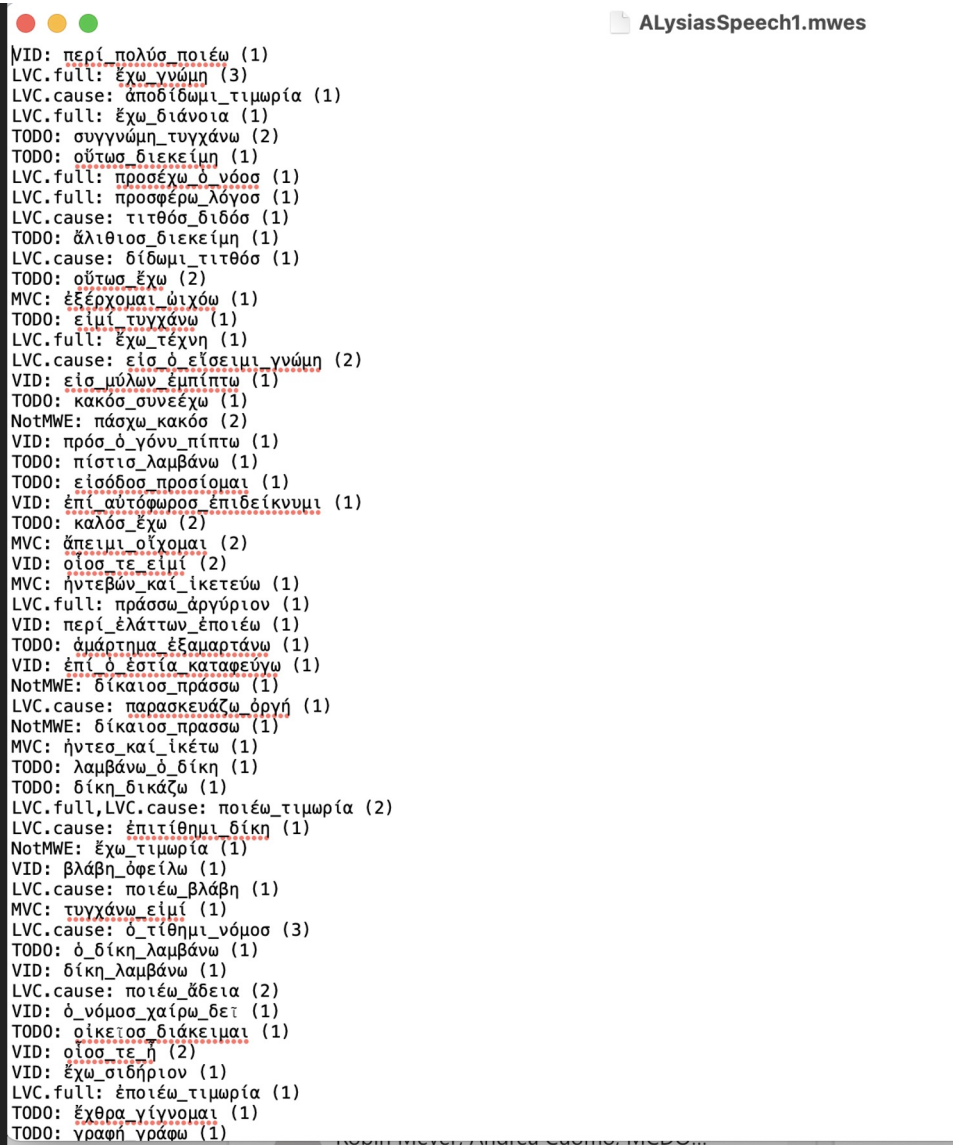
(1) [cf. VID] compositionality is categorial (see similarly Mel'cuk 2023)

Inter-annotator scores and the ‘Gold’ standard

https://gitlab.com/parseme/utilities/-/tree/master/st-organizers/corpus-statistics?ref_type=heads (extract_and_count_vmwes.py)

→ evaluate.py

→ consistencyCheckWebpage.py (comparison)



*Can VMWEs tell us something about ‘style’?

Comments on Lysias 1:

Carey (1989, p. 8) ‘The dominant impression created is one of artlessness’ (about Lysias)

Van Emde Boas (2022) describes Lysias’ creation Euphiletus as a ‘simple, homely man’.


UNIVERSITY OF OXFORD
FACULTY OF CLASSICS
LEVERHULME TRUST
Interfacing: The PARSEME Ancient Greek Corpus
A hands-on workshop
Churchill College, Cambridge
11 March 2024
Let's take a BREAK !

Programme

9:00–10:30am Session 1 (Annotation guidelines for a corpus language)

10:30–11:00 Break

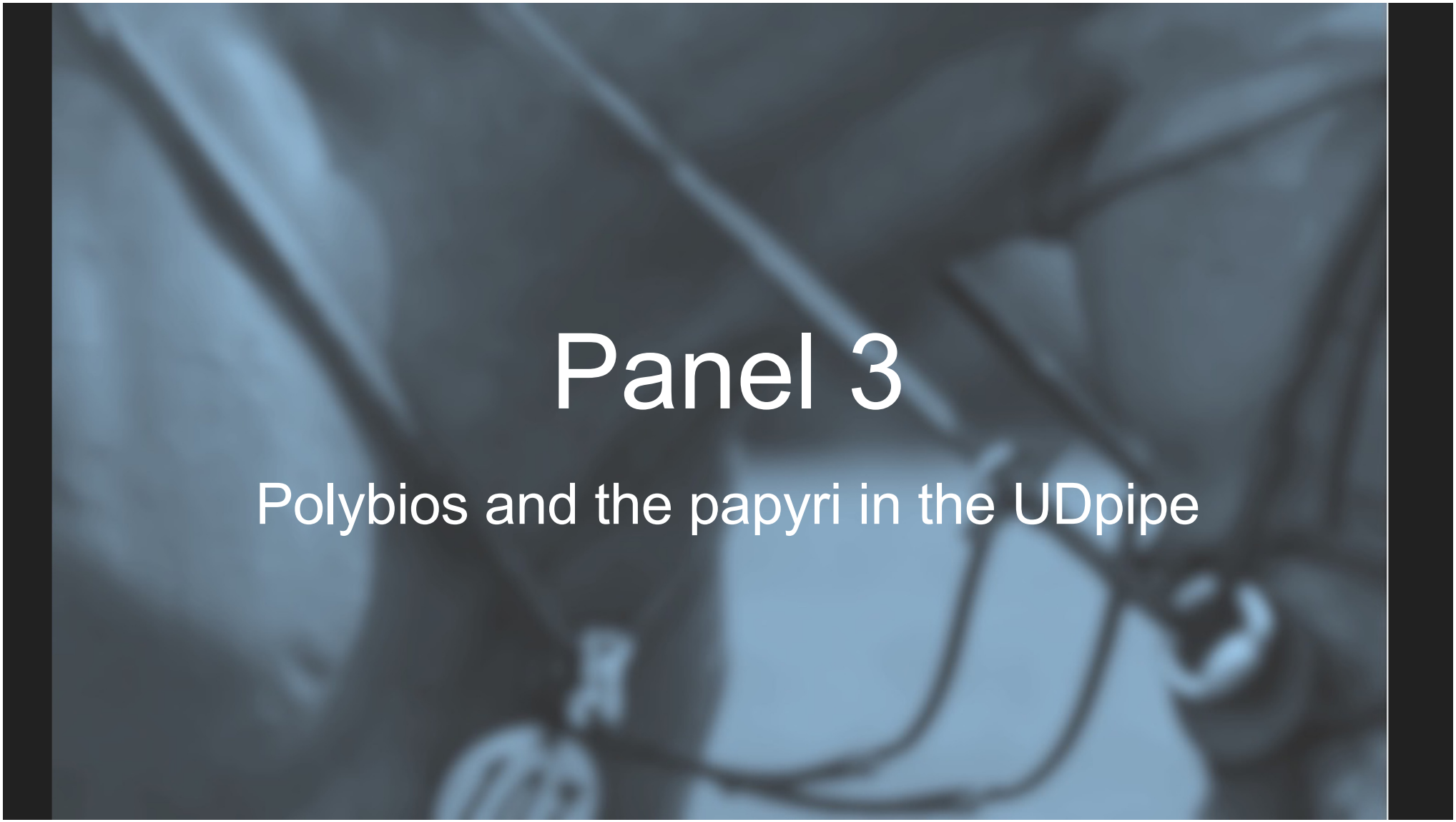
11:00–12:30 Session 2 (Literary classical Attic and inter-annotator scores)

12:30–2:00pm Sandwich lunch (hall)

2:00–3:30pm Session 3 (Polybios and the papyri in UD pipe)

3:30–4:00pm Break

4:00–5:30pm Session 4 (Wikipedia Light-verb constructions & nominal multi-word constructions)


Panel 3
Polybios and the papyri in the UDpipe

Post-classical Greek

Text 1: P. Kell. Gr. 1 68 & P. Neph. 9

- Problem 1: non-standard orthography & Co. → https://papygreek.com/text/437327 & https://papygreek.com/text/442195

Text 2: Polybios, Histories, book 1, chapter 1

- Problem 2: diachronic development of the language → the models are built for too internally diverse a corpus

Document for results: https://docs.google.com/document/d/1XfKVYTpld_ulDhQ-ECd0YWdhXye1PhywkJoxFx6r_Ik/edit?usp=sharing (also in your booklet and in email)

Emergence of new categories (VPC and IAV)

Phrasal verbs

- Homer's tmesis
- αἴρω ἀπό (IAV) & αἴρω ἔξω (VPC)

(Fendel 2020)

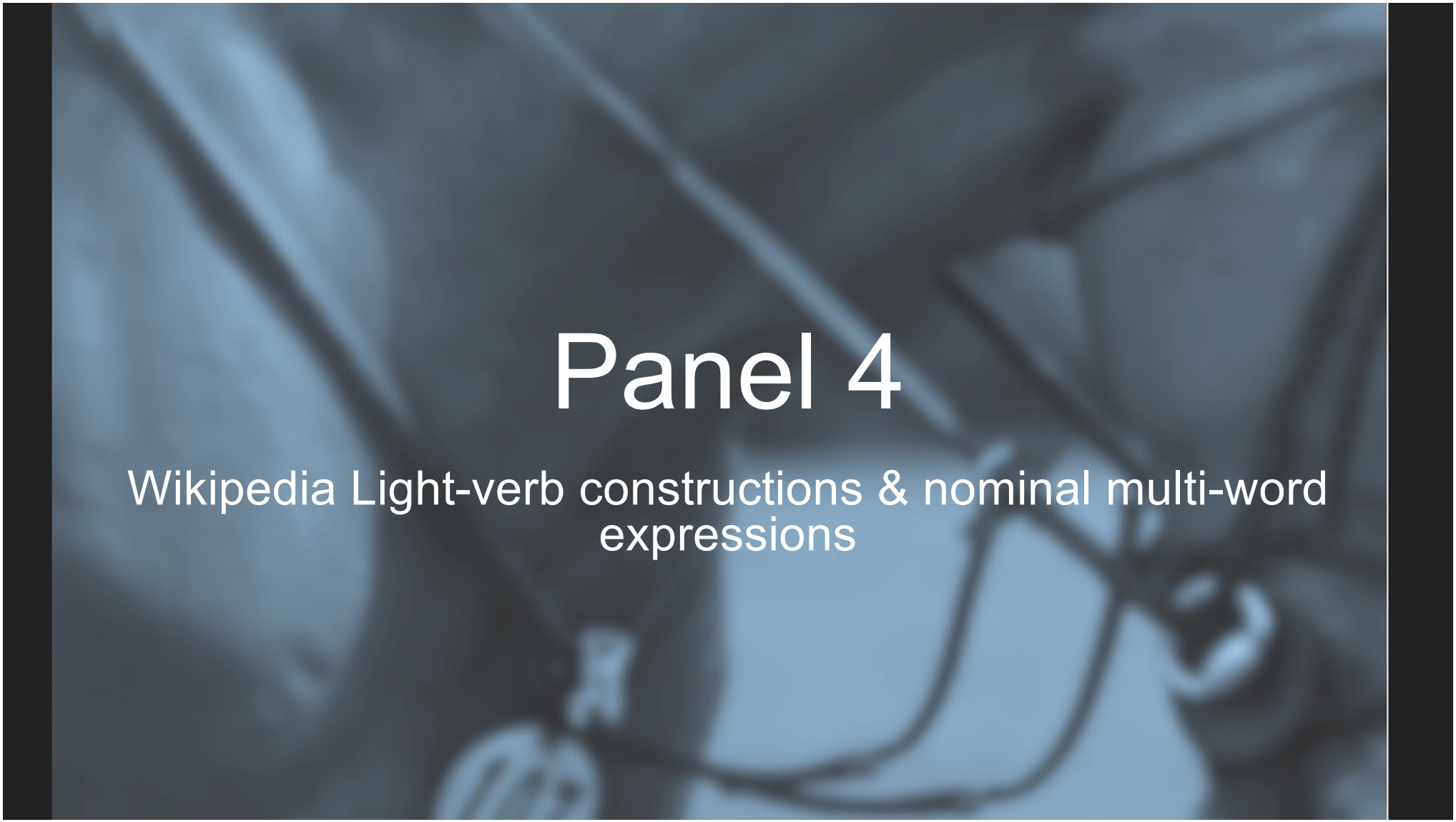
Section 5.5

VPC-specific decision tree:

↳Apply test VPC.1 - [**PART-REDUC**: *Can the verb without the particle refer to the same event?*]
↳ **NO ⇒** It is a **VPC.full**.
↳ **YES ⇒** Apply test VPC.2 - [**PART-SPATIAL**: *Is the particle spatial?*]
↳ **YES ⇒** It is **not** a VPC, exit
↳ **NO ⇒** Apply test VPC.3 - [**PART-SPATIAL-LIT**: *Is the particle spatial in a literal reading?*]
↳ **NO ⇒** It is a **VPC.semi**
↳ **YES ⇒** It is **not** a VPC, exit

- In fully non-compositional VPC (**VPC.full**) the change in the meaning of *v* goes significantly beyond adding the meaning of *p*:
  5.5_B_vpc-full to ***do in***
- In semi-non-compositional VPCs (**VPC.semi**), *p* adds a partly predictable but non-spatial meaning to *v*
  5.5_C_vpc-semi to ***eat up***

Note that in this shared task we *do not account for compositional verb-particle combinations*, i.e. those whose meaning can be deduced from the meaning of the preposition and of the verb.


Panel 4
Wikipedia Light-verb constructions & nominal multi-word expressions

Wikipedia – shared resource for annotators

Wikipedia s.v. Light verb: https://en.wikipedia.org/wiki/Light_verb

(see also Multi-word expression; PARSEME currently missing)

LINK for collaborative document: https://docs.google.com/document/d/1XfKVYTpId_uIDhQ-ECd0YWdhXye1PhywkJoxFx6r_Ik/edit?usp=sharing (also in booklet and via email)

→ e.g. the Ancient Greek specific annotations guidelines (currently the TODO category), issues identified with our specific annotation practice (cf. corpus language), etc.

Wikipedia article

Haspelmath 2010, p. 665 ‘**Comparative concepts** are concepts created by comparative linguists for the specific purpose of crosslinguistic comparison. Unlike descriptive categories, they are not part of particular language systems and are not needed by descriptive linguists or by speakers. They are not psychologically real, and they cannot be right or wrong. They can only be more or less well suited to the task of permitting crosslinguistic comparison. They are often labeled in the same way as descriptive categories, but they stand in a many-to-many relationship with them (§ 9). Comparative concepts are universally applicable, and they are defined on the basis of other universally applicable concepts: universal conceptual-semantic concepts, general formal concepts, and other comparative concepts. Comparative concepts have often been used implicitly in the typological literature, but there has not been any detailed and explicit discussion of the difference between comparative concepts and language-particular descriptive categories.’

Haspelmath 2010, p. 664
‘Each language has its own categories, and to describe a language, a linguist must create a set of **DESCRIPTIVE CATEGORIES** for it, and speakers must create mental categories during language acquisition. These categories are often similar across languages, but the similarities and differences between languages cannot be captured by equating categories across languages.’

Beyond verbal multi-word expressions (VMWEs)

Test DIST

Nominal MWEs

- NID
- VMWENom
- PronID (closed list)

Modifier MWEs

- AdjID
- AdvID

Functional MWEs

Generic decision tree

Details and **multilingual examples** of all tests are given in a dedicated document.

- Apply test DIST - [**DIST**: What is the distribution of the canonical form of the candidate c?]
  - Determiner, conjunction or adposition ⇒ Apply the guidelines for functional MWEs ⇒ *tests positive?*
    - Annotate with the category determined via the guidelines
    - It is not a functional MWE, **exit**
  - Adjectival or adverbial phrase ⇒ Apply the guidelines for modifier MWEs ⇒ *tests positive?*
    - Annotate with the category determined via the guidelines
    - It is not a modifier MWE, **exit**
  - Verb or Verbal phrase or verbal clause ⇒ Apply the VMWE guidelines ⇒ *VMWE tests positive?*
    - Annotate with the category determined via the guidelines
    - It is not a VMWE, **exit**
  - Noun or Nominal phrase ⇒ Apply the guidelines for nominal MWEs (below) ⇒ *tests positive?*
    - Annotate with the category determined via the guidelines
    - It is not a NMWE, **exit**

VMWENom & Functional MWEs

VMWENom

Pl. Rep. 407b2 νοσοτροφία τεκτονικῇ μὲν καὶ ταῖς ἄλλαις τέχναις ἐμπόδιον τῇ προσέξει τοῦ νοῦ ‘nursing a disease is a hindrance to the paying attention to carpentry and the other arts’

Functional MWEs

- Determiner ???
- Conjunction Coptic ⲉⲃⲟⲗ-ϫⲉ ADV-CONJ ‘because’ & ⲉⲧⲃⲉ-ϫⲉ PP-CONJ ‘because’
- Adposition: EBG papyri εἰς λόγονGEN ‘on account of’ (LSJ s.v. λόγος I.2) (cf. Bortone 2010 pp. 252–253; see also Matushansky and Zwarts 2021; Hoffmann 2005, ch. 8)

Next steps

PARSEME Ancient Greek working group

- Annotation and release (in autumn)
- Workshops for students to train them up (TORCH?)
- Publication (towards the end of the year)

CfP

- See next slide!

Wikipedia article

- A resource to share disagreements

CfP

[Papy] CfP Machine Learning for Ancient Languages, Bangkok Aug. 15

Isabelle Santaniello via PAPY <papy@lists.hum.ku.dk>
Wed 21/02/2024 18:16
To:papy@lists.hum.ku.dk <papy@lists.hum.ku.dk>
Cc:Isabelle Santaniello <imarthot@yahoo.com>

1 attachments (458 bytes)
ATT00001.txt;

+ apologies for cross posting +

On behalf of the organising Committee

Dear Papy-list Members,

On behalf of the Organising Committee, it is my pleasure to circulate the 1st Call for Papers for the Workshop on Machine Learning for Ancient Languages (ML4AL 2024).
The Workshop is co-located at ACL 2024 and will take place in a hybrid format in Bangkok, Thailand and remotely, on 15 August 2024.

The submission deadline is **May 17th**, 2024 11:59pm, UTC-12 (anywhere on Earth).
Please refer to the ML4AL workshop website http://ml4al.com for the full CfP and for more information.


UNIVERSITY OF OXFORD
FACULTY OF CLASSICS
LEVERHULME TRUST
Interfacing: The PARSEME Ancient Greek Corpus
A hands-on workshop
Churchill College, Cambridge
11 March 2024
THANK YOU VERY MUCH for joining, contributing, and not least annotating!