Research report: Annotation guidelines for multi-word expressions in corpus languages (with focus on classical Greek and Egyptian), Short-Term Scientific Mission, University of Jaén, Spain, 31/03/2025–14/04/2025

Funding: This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

Corpus languages in the PARSEME 2.0 guidelines: Challenges and workable solutions

Victoria B. Fendel
University of Oxford, UK
https://orcid.org/0000-0001-6302-3726

1 Resources (corpora and tools)

Greek is comparatively speaking a well-resourced corpus language, at least literary classical Attic, the variety of Greek that forms the basis for the PARSEME GRC corpus. Classical Greek, of the fifth and fourth centuries BC, is diatopically and diastratically diverse which makes it necessary to define a variety to focus on. This diversity is regularly reflected in multi-word expressions (Fendel 2025; Fendel 2024a). Classical literary Attic (dating to the fifth and fourth centuries BC and attested primarily in Athens) is traditionally considered the standard variety.

The corpus selection was done in line especially with the text types that form part of the modern Greek PARSEME corpus, including parliamentary speeches and newspaper reports. The ancient Greek corpus thus contains courtroom speeches, which neither abound with technical terminology nor with specific phraseology primarily because Athens in the classical period was a direct democracy. This necessitated that courtroom speeches were commonly comprehensible (Willi 2003: chap. 3). The corpus further includes dialogic passages about the constitution of a state, which due to their dialogic character retain an element of natural language use. Finally, the corpus contains the relatively unadorned report of a Greek mercenary army returning to Greece after a campaign in Persia.

The texts of the corpus have been digitised by the Perseus project and form part of the Perseus treebank. All markup related to textual criticism (incl. all Leiden markup) has been removed from the texts. This makes it necessary to use the corpus in parallel with a critical edition and was done for computational reasons not for linguistic ones. As our native speakers

are the texts (Fleischman 2000) and as grammars (and dictionaries) are written descriptively deducing rules (and patterns) from the extant corpus, the decision was made not to rely on the generation of forms based on the rules set out in grammars but to rely on attested structures only. Standard dictionaries of Greek (esp. Liddell-Scott-Jones) do not cover multi-word expressions comprehensively but only commonly used idiomatic ones (Fendel 2024b). Attestations can be searched in the full corpus of classical literary Attic non-verse texts for instance by means of the *Thesaurus Linguae Graecae* or the *Diorisis* tool (the process is discussed in detail in Fendel et al. (submitted)). Both tools allow for lemma-based searches within a specified distance of each other. The evaluation of the search results is manual. The texts are annotated by a team of four without annotators ever having vision of each other's annotation, such that inter-annotator scores can be calculated.

2 Findability (cranberry words and pronouns) [issue raised on Gitlab]

A cranberry item is "a token that does not have the status of a stand-alone word, has no proper distribution, and no stand-alone meaning, but it may have a syntactic category and an inflection paradigm. It only occurs in a particular expression (or a closed list of expressions) and can never be found in different contexts" according to the PARSEME 2.0 guidelines. Cranberry items are difficult to discover and identify because (i) we have to rely on negative evidence, i.e. an item has been checked in all the extant texts and is indeed distributed like this, (ii) we run the risk of over-generating cranberry items when focussing on smaller sub-samples.

Given PARSEME's fiercely synchronic focus, one source of candidates for cranberry items are so-called archaisms, i.e. features that have been revived in a language variety although having died out before without continuous attestation, and archaic features, i.e. features that show continuous attestation from the previous stage of the language into the one of interest but that are clearly decreasing in use, as shown by both token frequency and the number and type of contexts in which the item appears (cf. reallocation). The issues of negative evidence and the risk of overgeneration mentioned above exist for these too (for a vivid example of this, see the iconic debate on tmesis as an archaism vs. archaic feature) (e.g. Hajnal 2004 for references).[1] Categories of items that tend to preserve archaic features are onomastics (places names and personal names) and epithets. Another candidate category for cranberry

[1] Tmetic constructions would qualify as (Idiomatic) Verb Particle Constructions under the PARSEME 2.0 guidelines. This is not relevant to the current annotation campaign as they do not appear in the literary classical Attic non-verse texts of the GRC corpus.

items are suppletive paradigms, i.e. a paradigm in which an item is inserted in a slot where the historically derived item has fallen out of use.

3 Poetic licence (testing of attestations in corpora)

Especially a poet's stylistic liberties have been labelled a poetic licence since ancient times (e.g. already in Isocrates, fifth / fourth centuries BC, Nünlist (2009: chap. 7)). The term stylistic is rather vague and a concept that is more descriptive would be that of idiolect, i.e. one person's preferences with regard to language use which may or may not align with the preferences of the group of people the language user is part of (e.g. Biber & Conrad 2009). Idiosyncratic features can appear for a plethora of reasons ranging from humoristic intent, e.g. transformation of an idiomatic expression for humoristic purposes (cf. Plaza 2006; Savary et al. 2019), to the influence of bilingualism and language contact (e.g. Fendel 2022), among others.

From a corpus-linguistic perspective, this means that we need to make allowance for such idiosyncratic variation alongside e.g. diatopic and distratic variation. Sheinfux et al. (2019) observe that the larger the corpus the more likely such variation in the form of one-off instances is visible. Thus, for corpus languages such as classical Greek, the testing of an expression e.g. for lexical and syntactic inflexibility needs to take into account that a small percentage of variation does not necessarily mean that a structure is lexically and syntactically flexible but merely that one or some individuals pushed the boundaries. For example, one can check for idiosyncratic usage, i.e. the pattern being linked to one individual only, or whether the same individual uses the varied pattern consistently but also for where the observed variation clusters as regards e.g. the genre of text.

4 Derivational & invariable items (language-internal system) [issue raised on Gitlab]

The enrichment of the universal guidelines from the focus on verbal multi-word expressions to a focus on also functional, adjectival and adverbial, and nominal multi-word expressions in the context of the PARSEME/UniDive Shared Task 2025 has raised the issue of whether deverbal categories have to be derived lexically (by means of derivational morphology) or can also be derived morpho-syntactically (by means of inflexional morphology).

Classical Greek has a range of derivational morphology to derive nouns from verbs (e.g. formations in -μα and -σι-). The derivational morphology is not fully productive in the sense of generality (Barðdal 2008) but constraints apply e.g. phonetically (cf. Gunkel 2011). Conversely, the conversion of a verbal form into a nominal form by morpho-syntactic means

is fully productive in the sense of generality. This kind of morpho-syntactic nominalisation is not always indicated in the surface representation.

A further issue is whether to consider participial forms as is done traditionally as a verbal mood and thus as verbal category. Participial forms are fully productive in the sense of generality in classical Greek and in this regard align with the inflexional morphology. Conversely, one could take the approach of considering them based on their distribution as is done in other places in PARSEME, i.e. predicative participles would then be a verbal category and attributive participles would be a nominal category.

Another part-of-speech related issue concerns pragmaticalised phrases such as *I mean* and *you know* in English which are distributed like discourse adverbials and function outside of the sentence grammar (e.g. Molinelli 2010; la Roi 2022). They are invariable on the outside, similarly to discourse particles. Given the synchronic focus of the PARSEME annotation, their diachronic pathway should not be considered as an indication of the category they distribute like in classical Greek.

5 Language-specificity (categories that exist and do not exist, typology) [issues raised on Gitlab]

The universal annotation guidelines that the PARSEME initiative has developed for multi-word expressions are in Haspelmath's (2010) terms a comparative concept, i.e. "concepts created by comparative linguists for the specific purpose of crosslinguistic comparison … [and] are not psychologically real". This comparative concept needs to be translated into a descriptive category, i.e. psychologically real "language-specific categories … [in order] to identify the phenomena of a language with these preestablished crosslinguistic categories", for each participating language and its tests. E.g. English word order is strict SVO and syntax-driven such that an SVO order is not very indicative of a syntactically inflexible phrase; classical Greek word order is information-structurally driven (e.g. Dik 1995; Celano 2013; Matić 2003) and the unmarked order seems to tend towards SOV such that contiguity of components is considerably more indicative.

In classical literary Attic Greek oratory, historiography, and philosophical prose, several categories of verbal multi-word expressions do not exist, including Inherently Reflexive Verbs (IRVs), Inherently Adpositional Verbs (IAVs), and Idiomatic Verb-Particle Constructions (IVPCs). In all cases, competing morphological and morpho-syntactic means seem to exist in the system, i.e. verb lability and preverbation, and no competition (and possible reallocation) seems to have arisen at this stage in the language (for later stages, see Fendel

(2020)). Classical Greek also has a fully functional compounding system in the domain of the lexicon – compounds are usually right-headed but a small number of left-headed compounds exist (Tribulato 2015).

Typologically speaking, classical Greek is an Indo-European language with nominative-accusative alignment, a reasonably tensed verbal system, and rich inflexional and derivational morphologies. Classical Greek is a pro-drop language, i.e. a pronominal subject does not have to be indicated in the surface-representation if it is salient, and the language has a system of null anaphora of objects when they are discursively salient (e.g. Joseph 1994; Luraghi 2003; Luraghi 2004). This causes difficulty for the annotation process as the annotation platform FLAT does not currently allow for annotation of null placeholders or for the annotation of structures across sentences boundaries. The non-indication of salient pronominal subjects and salient objects in the surface representation is not to be considered an instance of ellipsis, i.e. the marked omission of an item for stylistic reasons (cf. Lausberg, Orton & Anderson 1998), but is a regular systemic feature.

6 Language models, treebanks, and Gitlab validation

Treebanks and language models built for corpus languages have to rely on the extant corpora which are often comparatively speaking small and internally heterogenous, by comparison to currently spoken languages. The ancient Greek working group uses the Perseus UD treebank and model as it relies on as large a sample as possible and comprises the texts of interest, i.e. literary classical Attic Greek. The Perseus UD treebank has the drawback that multi-word expressions are not implemented, i.e. ExtPos is not implemented. This means that issues can arise at the level of syntactic parsing in particular. The annotated corpora have to pass a validation test suit in Gitlab of both the UD columns and the PARSEME (multi-word expression) columns in the output .cupt file in order to ensure consistency.

References

Barðdal, Jóhanna. 2008. *Productivity: Evidence from case and argument structure in Icelandic*. *cal.8*. Amsterdam: John Benjamins.

Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.

Celano, Giuseppe. 2013. Argument-focus and predicate-focus structure in Ancient Greek. *Studies in Language* 37(2). 241–266.

Dik, Helma. 1995. *Word Order in Ancient Greek: A Pragmatic Account of Word Order Variation in Herodotus*. Leiden; Boston: Brill.

Fendel, Victoria. 2020. Phrasal verbs in a corpus of Post-Classical Greek letters from Egypt. In Martti Leiwo, Marja Vierros & Sonia Dahlgren (eds.), *Papers on Ancient Greek Linguistics Proceedings of the Ninth International Colloquium on Ancient Greek Linguistics (ICAGL 9) 30 August – 1 September 2018, Helsinki*, 63–98. Vaasa: Grano Oy.

Fendel, Victoria. 2022. *Coptic interference in the syntax of Greek letters from Egypt*. Oxford: Oxford University Press.

Fendel, Victoria. 2024a. Celebrating diversity: The origins and pathways of three support-verb constructions. *Lexis*.

Fendel, Victoria. 2024b. Epilogue: Taking wing. In Victoria Fendel (ed.), *Support-verb constructions in the corpora of Greek: Between lexicon and grammar?*, 327–340. Berlin: Language Science Press.

Fendel, Victoria. 2025. Taking stock of Greek support-verb constructions: synchronic and diachronic variability in the documentary papyri. In Jesus de la Villa et al. (ed.), *Advances in Ancient Greek Linguistics*, 295–311. Berlin; Boston: Mouton De Gruyter.

Fendel, Victoria, Elena Squeri & Paraskevi Platanou. submitted. Let's square a circle: The PARSEME-GRC corpus of verbal multi-word expressions. *Journal of Greek Linguistics*.

Fleischman, Suzanne. 2000. Methodologies and ideologies in historical linguistics: On working with older languages. In Susan Herring, Pieter Reenen & Lene Schøsler (eds.), *Textual parameters in older languages*, 33–58. Amsterdam: John Benjamins.

Gunkel, Dieter. 2011. The emergence of foot structure as a factor in the formation of Greek verbal nouns in -μα(τ)-. *Münchener Studien zur Sprachwissenschaft* 65. 77–103.

Hajnal, Ivo. 2004. Die Tmesis bei Homer und auf den mykenischen Linear B- Tafeln – ein chronologisches Paradox? In John Penney (ed.), *Indo-European Perspectives: Studies In Honour of Anna Morpurgo Davies*, 146–178. Oxford: Oxford University Press.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687.

Joseph, Brian. 1994. On weak subjects and pro-drop in Greek. In Irene Philippaki-Warburton, Katerina Nicolaidis & Maria Sifianou (eds.), *Themes in Greek Linguistics: Papers from the first international conference on Greek linguistics, September 1993*, 21–32. Amsterdam: John Benjamins.

Lausberg, Heinrich, David Orton & R. Dean Anderson. 1998. *Handbook of literary rhetoric: a foundation for literary study*. Leiden: Brill.

Luraghi, Silvia. 2003. Definite referential null objects in Ancient Greek. *Indogermanische Forschungen* 108. 167–194.

Luraghi, Silvia. 2004. Null objects in Latin and Greek and the relevance of linguistic typology for language reconstruction. In *Proceedings of the fifteenth annual UCLA Indo-European conference, Los Angeles, November 7–8, 2003*, 234–256. Washington, D.C.: Institute for the Study of Man.

Matić, Dejan. 2003. Topic, focus, and discourse structure: Ancient Greek Word Order: *Studies in Language* 27(3). 573–633.

Molinelli, Piera. 2010. From verbs to interactional discourse markers: the pragmaticalization of Latin rogo, quaeso. *Journal of Latin Linguistics* 11. 181–192.

Nünlist, René. 2009. *The Ancient Critic at Work: Terms and Concepts of Literary Criticism in Greek Scholia*. Cambridge: Cambridge University Press.

Plaza, Maria. 2006. *The Function of Humour in Roman Verse Satire: Laughing and Lying*. Oxford: Oxford University Press.

Roi, Ezra la. 2022. Weaving together the diverse threads of category change: Intersubjective ἀμέλει 'of course' and imperative particles in Ancient Greek. *Diachronica* 39(2). 159–192.

Savary, Agata, Silvio Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta & Voula Giouli. 2019. Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir. *The Prague Bulletin of Mathematical Linguistics* 112. 5–54.

Sheinfux, Livnat, Tali Greshler, Nurit Melnik & Shuly Winter. 2019. Verbal multiword expressions: Idiomaticity and flexibility. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions*, 35–68. Berlin: Language Science Press.

Tribulato, Olga. 2015. *Ancient Greek verb-initial compounds: their diachronic development withinthe Greek compound system*. Berlin; Boston: Mouton De Gruyter.

Willi, Andreas. 2003. *The languages of Aristophanes: aspects of linguistic variation in Classical Attic Greek*. Oxford: Oxford University Press.

Web resources

- UniDive CA21167: https://unidive.lisn.upsaclay.fr/doku.php?id=spin-off_and_related_national_projects

- GRC Working group: www.ancientgreekmwe.com
- Universal guidelines PARSEME: https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/index.php?page=053_Tests_for_ADJECTIVAL_and_ADVERBIAL_MWEs
- Gitlab PARSEME Shared Task 2025: https://gitlab.com/parseme/sharedtask-guidelines/-/issues