Identification and annotation of MWEs in corpus languages (Oxford, 10 July 2026)

Ioannou Centre for Classical and Byzantine Studies, University of Oxford, room 14/30 & Teams.

Programme

10:00 am	Welcome and project overview (Victoria B. Fendel)
10:30 am	Sanskrit (Andrea Farina & Alessandro Giudice) [in-person/online]
11:10 am	COFFEE BREAK
11:30 am	Greek (Elena Squeri & Irene Vlastou) [online]
12: 10 pm	Latin (Victoria B. Fendel) [in-person]
12:50 pm	LUNCH BREAK
14:30 pm	Armenian (Petr Kocharov) [online]
15:10 pm	Synergy 1: DiCoLat/DiCoGra (Lucía Madrigal Acero) [online]
15:40 pm	COFFEE BREAK
16:00 pm	Synergy 2: Universal Dependencies for Hittite (Maria Molina) [online]
16:30 pm	Synergy 3: Automatic detection of multi word expressions in Ancient Greek scientific/technical (mathematical) texts (Martijn Windelinckx) [in person]
17:00 pm	Closing remarks and final discussion (Victoria B. Fendel)
17:30 pm	END OF EVENT

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

Abstracts

Sanskrit

MultiWord Expression (MWE) annotation frameworks have been predominantly developed based on languages in which lexical meaning is distributed across discrete orthographic units. Sanskrit, by contrast, is characterised by a highly productive system of compounding (samāsa), which frequently encodes what would otherwise surface as multiword units in other languages. This structural property has significant consequences for both the identification and annotation of MWEs, calling for a re-evaluation of standard cross-linguistic criteria. This talk reports on annotation work conducted on selected prose passages from the Vedic Kauśikasūtra and the Sanskrit Pañcatantra. We highlight the methodological and theoretical challenges encountered, focusing in particular on the interaction between compounding and MWEs. A key observation is the relative scarcity of MWEs, which can be largely attributed to the functional load of compounds: many expressions that would qualify as MWEs cross-linguistically are instead lexicalised as single compound forms. As a result, Sanskrit MWEs tend to be restricted to a limited set of categories. The most frequent class is that of nominal idioms. Their predominance can be explained by the fact that compounds, while pervasive, do not fully eliminate nominal fixed expressions, especially in cases involving lexicalisation, semantic opacity, or conventionalised reference. On the other hand, verbal MWEs are comparatively rare. A comparison between the Vedic Kauśikasūtra and the Sanskrit Pañcatantra reveals a lower proportional presence of MWEs in the latter, likely reflecting diachronic and stylistic factors, including archaic syntax and a stronger reliance on inherited formulaic expressions that do not always align with standard MWE categories.

Greek: (Corpus) Language-internal issues in the identification of non-verbal multi-word expressions: some examples from Classical Greek

This paper examines some mismatches between the PARSEME 2.0 guidelines for annotating non-verbal multiword expressions (MWEs) and the morphosyntactic properties of Classical Greek (CG). All the examples insist on the difficulties in the categorisation of MWEs.

First, the distinction between verbal and nominal heads, telling verbal MWEs apart from non-verbal ones, is hard in languages allowing forms of syntactic nominalisation of non-finite verbal forms, like participles (Hopper & Thompson 1984; Croft 1991 §2.3; 2022 §2.2.3; Cooper 1998 § 50.4; Fendel 2025). The intermediate status of the participle, moreover, raises some issues in verb phrases involving *γίγνομαι* in genitive absolute constructions (*νυκτός/μάχης γενομένης*). These cases can be interpreted as both Deverbal adverbial idioms and Light verb constructions.

The capacity displayed by CG of nominalising full syntagms via the definite article (Guardiano 2013) also challenges the consistent annotation of adverbial idioms (*ὡς ἐπὶ τὸ πολύ*), whose

syntactic category may shift to adjectival or nominal (τὸ ὡς ἐπὶ τὸ πολύ). The polysemy of CG prepositions (Luraghi 2003: 75-81, Bortone 2010: 41-75) also causes prepositional phrases (κατά θάλατταν) to display varying degrees of irregularity and function either adverbially or adjectivally according to the context.

Segmentation difficulties arise in MWEs formerly annotated as Verbal idioms (περὶ πολλοῦ ποιοῦμαι «make a big deal out of...»), whose components can also be analysed in an adverbial MWE (e.g. περὶ πολλοῦ) combined with a standard verbal head. This is suggested by the combining of the adverbial sequence with different verbal heads (e.g. περὶ πολλοῦ ἐστιν «it is a big deal»). Problems of identification of lexicalised sequences are typical of corpus languages, for which we lack L1 speakers (Fleischman 2000). Finally, particle and participle constructions such as ὡς/καίπερ + participle (ὡς οἰκιοῦντες τὰς τότε καλουμένας Ἐννέα ὁδούς, καίπερ εἰδὼς τοὺς ἀνθρώπους), raise difficulties at segmentation (are the particles simply modifiers?), categorisation, and semantic ambiguity conditioned by the morphology of the participle. The ὡς construction is highly productive and largely compositional, which argues against the annotation as an Adverbial idiom, yet leaving it unannotated is unsatisfactory. καίπερ + participle shares the segmentation problem but differs in exhibiting a fixed concessive meaning, making it a stronger candidate for an annotation as Adverbial idiom and illustrating that categorisation within this construction type depends on the degree of semantic fixedness of the particle involved.

References

Bortone, P. 2010. Greek Prepositions: From Antiquity to the Present. Oxford.

Cooper, G. L. 1998. Attic Greek Prose Syntax. 1–2. Ann Arbor.

Croft, W. 1991. Syntactic Categories and Grammatical Relations. The Cognitive Organization of Information. Chicago

Croft, W. 2022. Morphosyntax. Constructions of the World's Languages. Cambridge–New York.

Fendel, V. 2025. Do Good or Suffer Evil: Syntactic Nominalisations in Verbal Multi-Word Expressions in the PARSEME GRC Corpus. Digital Studies in Language and Literature 2(1). 85-121.

Fleischman, S. 2000. Methodologies and ideologies in historical linguistics: On working with older languages. In S. Herring, P. Reenen & L. Schøsler (eds.), Textual parameters in older languages. Amsterdam. 33–58.

Guardiano, C. 2013. The Greek Definite Article Across Time. Studies in Greek linguistics 33. 76–91.

Haspelmath, M. 2010. Comparative Concepts and Descriptive Categories in Crosslinguistic Studies. Language 86(3). 663-687.

Hopper, P. J. & Thompson, S. A. 1984. The Discourse Basis for Lexical Categories in Universal Grammar. Language 60(4). 703–752

Luraghi, S. 2003. On the Meaning of Prepositions and Cases: The Expression of Semantic Roles in Ancient Greek. Amsterdam–Philadelphia.
PARSEME Guidelines 1.3 = Savary, A. et al. 2018. PARSEME Multilingual Annotation Guidelines. Version 1.3. (https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/)
PARSEME Guidelines 2.0 = Savary, A. et al. (υπό εξέλιξη). PARSEME Annotation Guidelines. Version 2.0. (https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/)

Armenian: Light verb constructions in Classical Armenian

The talk is dedicated to the syntactic and semantic properties of light verb constructions (LVCs) in Classical Armenian. I will introduce the types of LVCs represented in the MWE dataset and discuss the status of stative, inchoative, and factitive constructions formed with the verbs 'be', 'become' and 'make', respectively. Special attention will be paid to word-order constraints and the use of definiteness markers characteristic of shared arguments of transitive and ditransitive LVCs.

Latin: Verbal, adverbial, and conjunctional MWEs in Caesar and Tacitus

From Auden's (1894) phrase book to Flobert's (1966) seminal study of Latin support-verb expressions, verbal multi-word expressions have long been a noticeable trait of the Latin lexico-grammar, even if embedded in various discourses (e.g. colloquialism, periphrastic, idiom). Yet, beyond these, Latin has a wide range of conjunctional MWEs and adverbial MWEs at different stages of fixicity (e.g. *properea quod* vs. *post quam* vs. *postquam* along with Varro's *ex quis*). In the context of sub-task 1.6, 50% of the annotated data came from Caesar's *Gallic War* and 50% came from Tacitus' *Histories*. The contribution will show the adaptation of the universal annotation guidelines for MWEs to Latin and issues encountered, the distribution of types of MWEs vis-à-vis the language system of Latin, and specifically gradual univerbation of comparatively speaking high-frequency MWEs as opposed to the flexibility of MWEs in the famous long tail of the distribution.

Synergy 1: A proposal for the annotation of the lexical properties of LVCs in Ancient Greek: DiCoGra and REGLA vis à vis the PARSEME Greek corpus

Automated annotation of VMWEs in Ancient Greek has seen great progress over the past few years within the PARSEME project (Savary et al. 2015, 2026). While PARSEME has created digital tools for the detection and syntactic description of VMWEs, the Dictionary of Ancient Greek Collocations or DiCoGra (Jiménez López and Baños, n.d.) has made a solid theoretical and practical proposal for the analysis of verb-noun collocations based on the Meaning-Text Theory

(Mel'čuk 1997, 2023; Alonso Ramos 2004a, 2004b). The DiCoGra focuses on the lexical, syntactic and semantic description of LVCs, which is done manually by a team of scholars.

In this presentation, I argue that the incorporation of a layer of lexical annotation can contribute to a better understanding of the variety of constructions subsumed under the label of LVCs throughout the PARSEME corpora. While this would require manual annotation, especially in its initial stages, it has already been tested on other databases – remarkably, REGLA (Rección y Complementación del Griego y del Latín) – developed for the syntactic analysis of ancient languages (Garzón Fontalvo 2023). The main challenges in this regard concern (i) the semi-lexical nature of LVCs, (ii) the syntactic discontiguity of LVCs, and (iii) the semantic content of verb and noun – and their contribution to the overall meaning of the LVC.

The main objective of this presentation is to illustrate how the addition of this lexical layer could contribute to the organization of the data in PARSEME taking as an example some LVCs of the collocational pattern of ὀργή orgḗ 'rage' in Classical Greek. The Greek texts are retrieved from the TLG and will be made available online.

References

Alonso Ramos, Margarita. 2004a. 'Diccionario de colocaciones del español (DICE)'. http:/www.dicesp.com.

Alonso Ramos, Margarita. 2004b. Las construcciones con verbo de apoyo. Visor.

Garzón Fontalvo, Eveling, ed. 2023. Verba Notata: Retos de Anotación En Las Lenguas Clásicas. CSIC.

Jiménez López, María Dolores, and José Miguel Baños. n.d. 'DiCoGra: Diccionario de colocaciones del griego antiguo'. Accessed 20/05/2026. https://dicogra.iatext.ulpgc.es/.

Mel'čuk, Igor. 1997. 'Vers une linguistique sens-texte'. Collège de France, Chaire internationale.

Mel'čuk, Igor. 2023. General Phraseology: Theory and Practice. John Benjamins. https://doi.org/10.1075/lis.36.

Savary, Agata, Manfred Sailer, Yannick Parmentier, et al. 2015. 'PARSEME – PARSing and Multiword Expressions within a European Multilingual Network'. 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015) (Poznań, Poland).

Savary, Agata, Manon Scholivet, Carlos Ramisch, et al. 2026. 'PARSEME 2.0 Multilingual Corpus of Multiword Expressions'. In Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), edited by Stelios Piperidis, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral. European Language Resources Association (ELRA). https://doi.org/10.63317/2iy5qf38yhay.

Synergy 2: Universal Dependencies for Hittite

I am going to present preliminary results of our work on Universal Dependencies mark-up for Hittite letters. The letter of Puduheba, Queen of Hatti, to Ramesses II, Pharaoh of Egypt, serves as the primary case study. I will reveal grammar nuances that make it hard to follow standard UD guidelines and where a researcher needs a decision to create annotation both correct grammatically and complying with the demands of universality. A set of analytically nontrivial problems is discussed with examples: the treatment of direct speech marked clause by clause by =wa(r); the analysis of correlative constructions that do not align neatly with the adnominal relative-clause model presupposed by standard UD relation inventories; the segmentation of second-position clitic chains; the representation of mixed Hittite–Sumerian–Akkadian writing; and the representation of subjects in a partially pro-drop language. I will discuss the corpus itself, editorial principles, workflow, and annotation decisions that had been taken during the annotation of the Hittite material.

Synergy 3: Automatic detection of multi word expressions in Ancient Greek scientific/technical (mathematical) texts

Formulaicity is a hallmark of scientific language and terminology. This is a tendency to which Ancient Greek scientific language is no exception. This especially holds true for Ancient Greek 'deductive' sciences such as mathematics (Schironi, 2010 & 2014), in which field formulaic expressions can range from lower level multi word expressions such as the (elliptic) *ἡ AB (γραμμή) 'the line AB', to formulaic expression such as ἔστω X 'let there be (geometrical shape)', and to high-level comparison structures such as ὡς … οὕτως* 'as… so…'. Previous research on formulaicity in Ancient Greek has focused on epic formulaic language, but also on Ancient Greek Multi Word Expressions (MWEs) (Fendel 2025), as well as more specifically on formulaic expressions in Ancient Greek mathematical texts (Netz, 1999, 127-167). Still, this research was largely qualitative.

This presentation seeks to leverage quantitative methods to automatically detect candidates of formulaic expressions in Ancient Greek mathematical texts using the GLAUx corpus (Keersmaekers, 2021; 20 million tokens). We automatically select candidate formulaic expressions, not based on their linear proximity to each other in the sentence (as with N-grams), but rather on the basis of their syntactic relations. This is made possible by the syntactic annotation present in the GLAUx corpus. We then compute the significance of those co-occurrences with different measures. This methodology builds on the preexisting work of Martens & Vandeghinste (2010) and Fantoli & Korkiakangas (2025), while also considering preexisting tools for automatic MWE extraction like MWEtoolkit (Ramisch & Boitet, 2010; Ramisch 2020; Zagatti et al., 2022), and STARK (Krsnik & Dobrovoljc, 2025) for Universal Dependency treebanks

In a second step, we want to enrich these syntactic co-occurrences with semantic roles, which are also annotated in the GLAUx corpus. This accounts for semantic variation of constructions: it allows differentiation of syntactically equivalent constructions based on differences in the semantic roles of their constituents; conversely, it also facilitates detection of semantically equivalent but syntactically different constructions, such as is the case with passivization.

Finally, we plan to introduce automated provisional classification of the formula candidates: categories such as support verbs or light verb constructions can be identified by preselecting specific lemmas.

In adopting this exploratory strategy of formula detection, we reduce redundancy in two ways: as in previous research, we preclude interference of non-syntactic co-occurrences; next, we expand on this by exploiting semantic annotation, which prevents confusion of constructions on that score. The statistical information as well as the provisional categorization of said constructions serve as jumping off points for manual annotation. As such, our methodology facilitates the annotation and semantic classification of formulaic Ancient Greek expressions. In future this data may be used to train a language model to do the annotation automatically.

References

Fantoli, M., & Korkiakangas, T. (2025). Exploring formulaic language in dependency treebanks using network analysis. In Formulaic Language in Historical Linguistics: data, methods, tools, and theory, Date: 2025/06/02-2025/06/03, Location: University of Helsinki.

Fendel, V. B. (2025). Corpus languages in the PARSEME 2.0 guidelines: Challenges and workable solutions. https://zenodo.org/records/15198689

Keersmaekers, Alek. 2021. "The GLAUx corpus: methodological issues in designing a long-term, diverse, multi-layered corpus of Ancient Greek." Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021 , 39–50. Online: Association for Computational Linguistics. Available at: https://aclanthology.org/2021.lchange-1.6

Krsnik, L., & Dobrovoljc, K. (2025). STARK: A Toolkit for Dependency (Sub)Tree Extraction and Analysis. In S. Jablotschkin, S. Kübler, & H. Zinsmeister (Eds.), *Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025)* (pp. 44–51). Association for Computational Linguistics. https://aclanthology.org/2025.tlt-1.5/

Martens, S., & Vandeghinste, V. (2010). An Efficient, Generic Approach to Extracting Multi-Word Expressions from Dependency Trees.

Netz, R. (1999). The Shaping of Deduction in Greek Mathematics: A Study in Cognitive History (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511543296

Ramisch, C., Villavicencio, A., & Boitet, C. (2010, May). mwetoolkit: A framework for multiword expression identification. In *LREC* (Vol. 10, pp. 662-669).

Ramisch, C. (2020). Computational phraseology discovery in corpora with the mwetoolkit. In G. Corpas Pastor & J.-P. Colson (Eds.), *IVITRA Research in Linguistics and Literature* (Vol. 24, pp. 111–134). John Benjamins Publishing Company. https://doi.org/10.1075/ivitra.24.06ram

Schironi, F. (2014). Scientific Vocabulary. In Encyclopedia of Ancient Greek Language and Linguistics (pp. 262–265). Brill. https://doi.org/10.1163/2214-448X_eagll_COM_00000311

Schironi, F. (2010). Technical Languages: Science and Medicine. In A Companion to the Ancient Greek Language (pp. 338–353). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444317398.ch23

Zagatti, F., Medeiros, P. A. de L., Soares, E. da C., Silva, L. N. dos S., Ramisch, C., & Real, L. (2022). mwetoolkit-lib: Adaptation of the mwetoolkit as a Python Library and an Application to MWE-based Document Clustering. In A. Bhatia, P. Cook, S. Taslimipoor, M. Garcia, & C. Ramisch (Eds.), *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022* (pp. 112–117). European Language Resources Association. https://aclanthology.org/2022.mwe-1.16/

Speakers

Victoria Beatrix Fendel (D.Phil. Oxford, 2018) is a researcher in historical linguistics at the University of Oxford, editor of the Classics section of the Literary Encyclopaedia and the *Registered Reports in Linguistics*, and language leader for Ancient Greek in the PARSEME (UniDive COST) initiative (www.ancientgreekmwe.com). Her research focusses on language contact and multi-word expressions in literary, epigraphic, and papyrological sources. She is the editor of *Support-verb Constructions in the Corpora of Greek: Between Lexicon and Grammar?* (Language Science Press, 2024) and *Verbal Multi-Word Expressions in Corpus Languages: Data, Methodology, and Analysis* (Open Book Publishers, 2026) and the author of *Giving Gifts and Doing Favours: Literary Classical Attic Greek Support-verb Constructions* (Brill, 2025).

Andrea Farina is a Post-Doctoral Research Associate in Latin Linguistics in the Department of Digital Humanities, working on the ERC-selected project Computational Corpus Annotation for Quantitative Analysis of Latin Lexical Semantics (COALA). He holds a PhD in Digital Humanities (King's College London) and has a BA in Classics (University of Milan) and and MA in Theoretical and Applied Linguistics (University of Pavia). He specialises in syntactic and semantic annotation and has contributed to global projects such as the Greek, Latin, and Sanskrit WordNets and the Digital Corpus of Sanskrit. Andrea is committed to advancing open access research, particularly in Classical studies, and serves as Senior Editor for the Journal of Open Humanities Data. He is the creator and convenor of the Data in Historical Linguistics international seminar series. He is also group lead of King's Linguistics Network at King's College London.

Alessandro Giudice (Ph.D., University of Cagliari, 2026) is a postdoctoral researcher in Indology at Ludwig Maximilian University of Munich, where he is a member of the Cluster of Excellence "Cross-Cultural Philology" (EXC 3061/1). His research focuses on Sanskrit technical literature, particularly law (Dharmaśāstra) and grammar (Vyākaraṇa), as well as on the history of premodern South Asia, Gandhāra and the Gāndhārī language, and Nepal and the Classical Newar language. In addition to publishing several articles on these topics in international journals and conference proceedings, he is the co-author, with Tiziana Pontillo, of *Medhātithi's Grammatical Notes on the Mānavadharmaśāstra: How Technicalities Can Shed Light on Ancient Indian Law* (Corpus Iuris Sanscriticum et Fontes Iuris Asiae Meridianae et Centralis 11; Turin: Edizioni AIT–Edizioni ETS, 2025).

Elena Squeri (Ph.D. Università di Genova/Sorbonne Université, 2023) is a member of the working group for the annotation of Ancient Greek in the PARSEME (UniDive COST) initiative. Her research focusses on the morphosyntax-semantics interface in the Ancient Greek verbal system, on the birth and development of sectorial languages, and on the establishment of

grammatical and linguistic concepts in antiquity. She is the author of La specializzazione semantica nel sistema verbale del greco antico e nella lingua della ginecologia ippocratica (Schwabe, 2027).

Irene Vlastou is an annotator for the Ancient Greek (GRC) corpus of the PARSEME (UniDive COST) initiative. She holds a BA in Classics and an MA in papyrology. Her master's thesis, a palaeography handbook proposing Modern Greek translations for palaeographical terminology alongside a study of non-literary scripts in papyri, was awarded the Lysimachos Kaftantzoglou Prize for outstanding theses by the University of Athens. She is a co-author of "A Letter from Petesuchos to Ptolemaios from the Cologne Papyrus Collection" (ZPE 227, 2023). Her research interests lie in papyrology, Greek palaeography and digital humanities. She is currently pursuing a second master's degree in Digital Humanities at the Athens University of Economics and Business, focusing on natural language processing and handwritten text recognition.

Christopher

Petr Kocharov (PhD, Leiden University 2019; HDR, Inalco 2025) is a researcher specializing in historical linguistics and Armenian studies. His research focuses on the morphosyntax of Classical Armenian from comparative-historical and typological perspectives. He developed the Classical Armenian Valency Lexicon (CAVaL) research platform ([https://caval.dch.phil-fak.uni-koeln..de/]https://caval.dch.phil-fak.uni-koeln.de) and the UD Classical Armenian CAVaL treebank (https://universaldependencies.org/treebanks/xcl_caval), and is a co-editor of Studies in Armenian Grammar and Lexicon (Röll, 2024).

Lucía Madrigal Acero (MA Valladolid, 2022) is a predoctoral researcher at the Complutense University of Madrid. Her dissertation focusses on support-verb constructions with ποιέω in post-classical Greek. She has collaborated on the Dictionary of Latin Collocations poieō (https://dicolat.iatext.ulpgc.es/) and the Dictionary of Ancient Greek Collocations (https://dicogra.iatext.ulpgc.es/), coordinated by Baños and Jiménez López. She is a research associate for the project Interacción del léxico y la sintaxis en griego antiguo y latín 2: Diccionario de Colocaciones Latinas (DiCoLat) y Diccionario de Colocaciones del Griego Antiguo (DiCoGrA) (PID2021-125076NB-C42) funded by the Spanish Ministry of Science, Innovation and Universities (PRE2022/102525).

Maria Molina is a Hittitologist working on Hittite syntax and semantics with digital and corpus methods. She holds a Ph.D. from the Institute of Linguistics, Russian Academy of Sciences; her thesis, *Word order in Hittite: corpus methods and analysis in typological perspective*, was defended in 2019. In 2022 she moved to Israel and worked on the semantics of Hittite emotions and on Universal Dependencies for Hittite at Tel Aviv University under the supervision of Prof. Amir

Gilan. Since February 2026 she has been a lecturer in General Linguistics at the Polis Institute of Ancient Languages in Jerusalem.

Martijn Windelinckx is a PhD student at KU Leuven University. He studied Ancient Greek and Latin there at the faculty of arts from 2017 to 2020. In the two subsequent years, he also obtained his master's degree in Linguistics and Artificial Intelligence there. Afterwards he studied Dutch, English, and French. He is now in the second year of his PhD, studying the consistency of Ancient Greek and Latin scientific language, using corpus-based methods.