Journal of Open Humanities Data

Lost at Sea: A Dataset of 25+ SEA Words Morpho-Semantically Annotated in Ancient Greek and Latin

COLLECTION:
REPRESENTING THE ANCIENT WORLD THROUGH DATA

DATA PAPER

ANDREA FARINA

ubiquity press

ABSTRACT

This paper describes a dataset containing more than 25 Ancient Greek and Latin words (nouns, verbs, adjectives) connected to the semantic field SEA ('sea', 'water', 'wave', 'shore', 'sail', 'maritime'). Tokens have been morphologically and semantically annotated, distinguishing among literal, metaphorical, and metonymic senses according to cognitive linguistics. Data have been stored in Figshare and are publicly available. This dataset can serve as a model for cross-linguistic semantic analyses in this or in other semantic fields, not only in the languages considered here. It can also be used to retrieve information in other research areas, such as literature, geography, anthropology, and psychology.

CORRESPONDING AUTHOR:
Andrea Farina
Department of Digital Humanities, King's College London, London, United Kingdom
andrea.farina@kcl.ac.uk

KEYWORDS:
historical linguistics; cognitive linguistics; Ancient Greek and Latin semantics; semantic field SEA; metaphors

TO CITE THIS ARTICLE:
Farina, A. (2023). Lost at Sea: A Dataset of 25+ SEA Words Morpho-Semantically Annotated in Ancient Greek and Latin. *Journal of Open Humanities Data*, 9: 24, pp. 1–7. DOI: https://doi.org/10.5334/johd.139

(1) OVERVIEW

REPOSITORY LOCATION

https://doi.org/10.18742/23968773

CONTEXT

In linguistics, the semantic field SEA has been studied in different ways, from the semantics of verbs of navigation (e.g. Maisak & Rakhilina, 2007; Divjak et al., 2010; Lander et al., 2012 in linguistic typology; Farina, 2021 on Ancient Greek) to metaphors connected to the sea (e.g. Leotta & De Felice, forth. 2023 on Latin). In Greek and Roman culture, the sea holds a prominent position, militarily (Harris, 2017; Nash, 2018), economically (Reed, 2003; Wilkinson, 2020; Boardman et al., 2021), and culturally (Berens, 1979; Lindenlauf, 2004; Nikoloska, 2012; Beaulieu, 2016).

This dataset contains linguistic information about more than 25 nouns, verbs, and adjectives connected to the semantic field SEA in four Ancient Greek and Latin texts between 5th – 1st century BCE (Lat. *De Bello Gallico* by Caesar, *Aeneid 1–6* by Vergil; AGr. *Histories 1–2* by Herodotus, *Argonautica* by Apollonius Rhodius).

The dataset has been created to support research on how the concept of SEA is lexicalized in Ancient Greek and Latin poetry and prose, with a case study on four authors.[1]

(2) METHOD

In this section, I summarize the steps that I followed to obtain the dataset presented here.

STEPS

1. *Text retrieval*: after choosing the texts (see Section 1 and below), I downloaded them in .txt format from Perseus 5.0 – also called Scaife Viewer – of the Perseus Digital Library (Crane 1987; Crane et al. 2006).[2]
2. *Text annotation*: I then uploaded the texts on the annotation platform INCEpTION (Klie et al., 2018, then Boullosa et al., 2018; de Castilho et al., 2018a; de Castilho et al., 2018b; Klie, 2018; Klie et al., 2020), developed by the Ubiquitous Knowledge Processing (UKP) Lab at TU Darmstadt. I created my annotation tagsets and layers, based on the linguistic parameters that were of interest for my work, i.e. morphology, lemma, passage, semantics, meaning (literal, metaphorical, metonymic), relations with proper nouns (see Section 3 for a more detailed description). At the end of my annotation, I exported the data in the UIMA CAS XMI (XML 1.0) format.[3]
3. *Data extraction and dataset creation*: I used a Python script specifically designed for the UIMA framework to extract the annotated data. I created a dictionary based on token IDs where I mapped the annotation layers. I then exported the dataset resulting from this extraction in CSV format.

SAMPLING STRATEGY

For this dataset, I decided to focus on two literary genres, i.e. historiography (Lat. *De bello Gallico* by Caesar; Gr. *Histories 1–2* by Herodotus) and epic poetry (Lat. *Aeneid 1–6* by Vergil; Gr. *Argonautica* by Apollonius Rhodius). Given that I also wanted to investigate the distribution of SEA words in Ancient Greek and Latin, I selected parts of these texts depending on the final number of tokens. To maintain a balance between the Latin and Greek sub-corpora, some texts (Herodotus's *Histories* and Vergil's *Aeneid*) have not been fully annotated. Overall, my corpus has 174,501 tokens. The Greek sub-corpus constitutes 53% of the whole corpus, and it has 92,592 tokens (53,750 for prose and 38,842 for poetry). The Latin sub-corpus has 81,909 tokens (51,313 for prose and 30,596 for poetry).

1 The results of this study were presented at ICHL26, the International Colloquium of Historical Linguistics (Heidelberg, Germany, 4–8 September 2023), by Andrea Farina, William Michael Short, and Barbara McGillivray.

2 https://scaife.perseus.org (Last accessed: 27 October 2023).

3 https://uima.apache.org/d/uimaj-current/references.html#ugr.ref.xmi (Last accessed: 27 October 2023).

(3) DATASET DESCRIPTION

The nouns, verbs, and adjectives included in this dataset are:

- NOUNS: AGr. *thálassa*, *póntos*, *pélagos*, *háls*, Lat. *mare*, *pontus*, *pelagus*, *aequor* 'sea'; AGr. *húdōr*, Lat. *aqua*, *lympha* 'water'; AGr. *háls*, Lat. *sal* 'sea', 'salt'; AGr. *kûma*, Lat. *unda*, *fluctus* 'wave'; Lat. *litus*, *ripa* 'shore';[4]
- VERBS: AGr. *pléō* (and its preverbed forms occurring in the analyzed texts), Lat. *navigo* 'sail';[5]
- ADJECTIVES: AGr. *thalássios*, *póntios*, Lat. *marinus*, *maritimus* 'maritime, marine'.[6]

In the CSV file, annotations are represented with ten columns and as many rows as the number of SEA tokens in each of the considered texts. Columns provide: (1) the token (TOKEN); (2) its morphological analysis (MORPHOLOGICAL FEATURES); (3) its lemma (LEMMA); (4) its part of speech (POS); (5) the sentence in which the token is found (PASSAGE); (6) the type of token meaning (literal, metaphorical, or metonymic), according to cognitive linguistics and the new WordNets for ancient Indo-European languages (Biagetti et al. 2021) (MEANING); (7) its meaning in context using synsets from the WordNets, preceded by a unique identifier (SYNSET); (8) the token ID (ID); (9) possible words (proper nouns or adjectives) in Ancient Greek or Latin to which a noun meaning 'sea' is referred (REFERS TO); (10) the meaning of the phrase resulting from (1) and (9), using synsets from the WordNets, preceded by their unique identifier (DENOTES). An excerpt of the dataset is given in Table 1.

OBJECT NAME

25+ SEA words morpho-semantically annotated in Ancient Greek and Latin.

FORMAT NAMES AND VERSIONS

CSV

CREATION DATES

From 2023-07-07 to 2023-08-10

DATASET CREATORS

Andrea Farina (Department of Digital Humanities, King's College London): conceptualization, data curation, methodology, formal analysis, data retrieval.

LANGUAGE

Ancient Greek, Latin, English

LICENSE

CC0

REPOSITORY NAME

Figshare

PUBLICATION DATE

2023-08-18

(4) REUSE POTENTIAL

Given that this dataset describes the semantics of different words pertaining to the semantic field of SEA in Ancient Greek and Latin, its first reuse potential deals with linguistics. First, the

4 No occurrences for the AGr. counterpart *paralía* 'shore' were found in these texts.

5 No occurrences of preverbed forms of Lat. *navigo* were found in these texts.

6 No occurrences for Lat. *pelagius* 'maritime, marine' were found in these texts.

Table 1 An excerpt of the dataset (13 rows of Apollonius Rhodius's *Argonautica*).

TOKEN	MORPHOLOGICAL FEATURES	LEMMA	POS	PASSAGE	MEANING	SYNSET	ID	REFERS TO	DENOTES
ἀὰὸς	Case=Gen|Gender=Fem|Number=Sing	ἅλς	NOUN	ἔνθ' ἄρα τοίγε ἐσπέριοι ἀνέμοιο παλιμπνοίῃσιν ἔκελσαν, καί μιν κυδαίνοντες ὑπὸ κνέφας ἔντομα μήλων κεῖαν, ὁρινομένης ἀλὸς οἴδματι	Literal	'n#06781694 a large body of water constituting a principal part of the hydrosphere'	25434		
πόντῳ	Case=Dat|Gender=Masc|Number=Sing	πόντος	NOUN	ἠῶθεν δ' Ὁμόλην αὐτοςχεδὸν εἰςορόωντες πόντῳ κεκλιμένην παρεμέτρεον	Literal	'n#06781925 a division of an ocean or a large body of salt water partially enclosed by land'	25716		
ἀλὸς	Case=Gen|Gender=Fem|Number=Sing	ἅλς	NOUN	λάρνακι δ' ἐν κοίλῃ μιν ὕπερθ' ἀλὸς ἦκε φέρεςθαι, αἴ κε φύγῃ	Literal	'n#06781694 a large body of water constituting a principal part of the hydrosphere'	26911		
πόντον	Case=Acc|Gender=Masc|Number=Sing	πόντος	NOUN	ἀλλὰ γὰρ ἔμπης ἦ θαμὰ δὴ πάπταινον ἐπὶ πλατὺν ὄμμαςι πόντον δείματι λευγαλέῳ, ὁπότε Θρήικες ἴαςιν	Metonymic	'n#06783379 the part of the sea that can be seen from the shore'	27311		
ἀλὶ	Case=Dat|Gender=Fem|Number=Sing	ἅλς	NOUN	περὶ γὰρ βαθυλήιος ἄλλων νήςων, Αἰγαίῃ ὅςαι εἰν ἀλὶ ναιετάουςιν	Metonymic	'n#06781925 a division of an ocean or a large body of salt water partially enclosed by land'	36044	['Αἰγαίῃ']	['n#06806923 an arm of the Mediterranean between Greece and Turkey; a main trade route for the ancient civilizations of Crete and Greece and Rome and Persia']
ἀναπλώοντι	Case=Dat|Gender=Masc|Number=Sing|Tense=Pres|VerbForm=Part|Voice=Act	ἀναπλέω	VERB	εἰ δ' οὔ μοι πέπρωται ἐς Ἑλλάδα γαῖαν ἱκέςθαι τηλοῦ ἀναπλώοντι, ςὺ δ' ἄρςενα παῖδα τέκηαι	Literal	'v#01260993 travel by boat'	39277		
ὕδωρ	Case=Acc|Gender=Neut|Number=Sing	ὕδωρ	NOUN	ἔνθ' ἄρα τοίγε κόπτον ὕδωρ δολιχῇςιν ἐπικρατέως ἐλάτῃςιν	Literal	'n#10771040 water containing salts'	39681		
ἅλα	Case=Acc|Gender=Fem|Number=Sing	ἅλς	NOUN	ὅφρα δαέντες ἀρρήτους ἀγανῇςι τελεςφορίῃςι θέμιςτας ςωότεροι κρυόεςςαν ὑπεὶρ ἅλα ναυτίλλοιντο	Literal	'n#06781694 a large body of water constituting a principal part of the hydrosphere'	39864		
πόντου	Case=Gen|Gender=Masc|Number=Sing	πόντος	NOUN	κεῖθεν δ' εἰρεςίῃ Μέλανος διὰ βένθεα πόντου ἰέμενοι τῇ μὲν Θρῃκῶν χθόνα, τῇ δὲ περαίην Ἴμβρον ἔχον καθύπερθε	Literal	'n#06781925 a division of an ocean or a large body of salt water partially enclosed by land'	40063	['Μέλανος']	['n#06810637 a sea between Europe and Asia; a popular resort area of eastern Europeans']
πέλαγος	Case=Acc|Gender=Neut|Number=Sing	πέλαγος	NOUN	πέλαγος δὲ τὸ μὲν καθύπερθε λέλειπτο ἦρι	Literal	'n#06783080 an especially deep part of a sea or ocean'	40294		
ἅλα	Case=Acc|Gender=Fem|Number=Sing	ἅλς	NOUN	ἔςτι δέ τις αἰπεῖα Προποντίδος ἔνδοθι νῆςος τυτθὸν ἀπὸ Φρυγίης πολυληίου ἠπείροιο εἰς ἅλα κεκλιμένη	Literal	'n#06781694 a large body of water constituting a principal part of the hydrosphere'	40710		
ὕδατος	Case=Gen|Gender=Neut|Number=Sing	ὕδωρ	NOUN	ἐν δέ οἱ ἀκταὶ ἀμφίδυμοι, κεῖνται δ' ὑπὲρ ὕδατος Λίςήποιο	Metonymic	'n#06789983 a large natural stream of water (larger than a creek)'	40823		
ἀλός	Case=Gen|Gender=Fem|Number=Sing	ἅλς	NOUN	ἠοῖ δ' εἰςανέβαν μέγα Δίνδυμον, ὄφρα καὶ αὐτοὶ θηήςαιντο πόρους κείνης ἀλός	Literal	'n#06781925 a division of an ocean or a large body of salt water partially enclosed by land'	42805		

dataset can lead to both onomasiological and semasiological analyses. It can be expanded considering other works, authors, and literary genres, to have a broader overview of SEA words in Ancient Greek and Latin. Similar datasets may also be obtained for other semantic fields and/or languages, to allow for cross-linguistic comparisons either synchronically or diachronically. Moreover, this dataset could serve as the basis to train a model for automatic semantic annotation based on co-occurring words, that can be extracted from the passage in which a token occurs.

This dataset – or other similar datasets – may also be employed in literary-geographical studies, to evaluate, for instance, how a specific place, such as a sea, is referred to in different texts and/or geographical areas – synchronically or diachronically –, and whether the proper noun of a sea tends to occur alone or with one or more common nouns. This may cast some new light on geographical denominations in the ancient world. In this sense, it may also be used to expand already existing online resources, such as Pelagios[7] (Simon et al., 2012; Barker et al., 2016; Simon et al., 2016; Kahn et al., 2021; Vitale et al., 2021) or to add further historical depth to the World Historical Gazetteer[8] (Manning & Mostern, 2015; Manning, 2015; Mostern, 2017), grouping together places that were called with more than one name.

Finally, more broadly, cross-linguistic analyses conducted in a cognitive framework also allow for psycho-anthropological studies that can address questions such as: How many words did the Greeks and the Romans possess to express one or more concepts related to SEA? How and why does the number of SEA words vary in Greek and Roman texts? How can we account for similarities and differences in this sense? Does this reveal anything about these populations from the cultural point of view?

ACKNOWLEDGEMENTS

I would like to thank Dr Barbara McGillivray for her precious linguistic, computational, and stylistic feedback on this work. I also thank Paola Marongiu who read a preliminary version of the paper.

FUNDING INFORMATION

This dataset has been produced to present a paper at the International Colloquium of Historical Linguistics (see fn. 1), whose expenses were covered by the AHRC London Arts & Humanities Partnership.

COMPETING INTERESTS

I am guest editor of the special collection *Representing the Ancient World through Data* and social media manager of this journal and did not take part in the editorial process pertaining to this manuscript.

AUTHOR AFFILIATIONS


Andrea Farina orcid.org/0000-0002-1948-9008
Department of Digital Humanities, King's College London, London, United Kingdom

REFERENCES

Barker, E., Simon, R., Isaksen, L., & **de Soto Cañamares, P.** (2016). The Pleiades Gazetteer and the Pelagios Project. In M. L. Berman, R. Mostern, and H. Southall (Eds.), *Placing Names: Enriching and Integrating Gazetteers*, 97–109. Bloomington: Indiana University Press. DOI: https://doi.org/10.2307/j.ctt2005zq7.12

Beaulieu, M. C. (2016). The sea in the Greek imagination. Philadelphia: PENN/University of Pennsylvania Press. DOI: https://doi.org/10.9783/9780812291964

Berens, E. M. (1979). The myths and legends of ancient Greece and Rome. Boston: Longwood Press.

Biagetti, E., Zanchi, C., & **Short, W. M.** (2021). Toward the creation of WordNets for ancient Indo-European languages. *Proceedings of the 11th Global Wordnet Conference, University of South Africa (UNISA)*. Global WordNet Association, 258–266.

7 https://pelagios.org (Last accessed: 27 October 2023).

8 https://whgazetteer.org (Last accessed: 27 October 2023).

Boardman, J., Tsetskhladze, G. R., Avram, A., & **Hargrave, J.** (2021). *The Greeks and Romans in the Black Sea and the importance of the Pontic region for the Graeco-Roman world (7th century BC-5th century AD): 20 years on (1997–2017)*. Proceedings of the Sixth International Congress on Black Sea Antiquities (Constanta – 18–22 September 2017). Oxford: Archaeopress. DOI: https://doi.org/10.2307/j.ctv1pdrqhw

Boullosa, B., de Castilho, R. E., Kumar, N., Klie, J. C., & **Gurevych, I.** (2018). Integrating Knowledge-Supported Search into the INCEpTION Annotation Platform. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 127–32. DOI: https://doi.org/10.18653/v1/D18-2022

Crane, G. (1987). From the Old to the New: Integrating Hypertext into Traditional Scholarship. *HYPERTEXT '87: Proceedings of the ACM Conference on Hypertext*, 51–57. New York, NY, USA: ACM Press. DOI: https://doi.org/10.1145/317426.317432

Crane, G., Bamman, D., Cerrato, L., Jones, A., Mimno, D., & **Packel, A.** (2006). Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries. *Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006)*, Alicante (Spain), 341–52. DOI: https://doi.org/10.1007/11863878_30

de Castilho, R. E., Klie, J. C., Kumar, N., Boullosa, B., & **Gurevych, I.** (2018a). INCEpTION – Corpus-based Data Science from Scratch. *Digital Infrastructures for Research (DI4R) 2018*, 9–11 October 2018, Lisbon, Portugal.

de Castilho, R. E., Klie, J. C., Kumar, N., Boullosa, B., & **Gurevych, I.** (2018b). Linking Text and Knowledge using the INCEpTION annotation platform. *Proceedings of the 14th eScience IEEE International Conference*, Amsterdam, Netherlands. DOI: https://doi.org/10.1109/eScience.2018.00077

Divjak, D., Koptjevskaja-Tamm, M., & **Rakhilina, E. V.** (2010). Aquamotion verbs in Slavic and Germanic. A case study in lexical typology. In V. Hasko, R. Perelmutter (Eds.), *New Approaches to Slavic Verbs of Motion*. Amsterdam/Philadelphia: John Benjamins, 315–341. DOI: https://doi.org/10.1075/slcs.115.18kop

Farina, A. (2021). *Aquamotion Verbs in Ancient Greek. A Study on pléō and its Compounds* (MA dissertation, University of Pavia).

Harris, W. V. (2017). Rome at sea: the beginnings of Roman naval power. *Greece & Rome*, *64*(1), 14–26. DOI: https://doi.org/10.1017/S0017383516000218

Kahn, R., Leif, I., Barker, E., Simon, R., de Soto, P., & **Vitale, V.** (2021). Pelagios – Connecting Histories of Place. Part II: From Community to Association. *International Journal of Humanities and Arts Computing*, *15*(1–2), 85–100. DOI: https://doi.org/10.3366/ijhac.2021.0263

Klie, J. C. (2018). INCEpTION: Interactive Machine-assisted Annotation. *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems (DESIRES)*, Bertinoro, Italy, 105.

Klie, J. C., Bugert, M., Boullosa, B., de Castilho, R. E., & **Gurevych, I.** (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING 2018)*, Santa Fe, New Mexico, USA, 5–9. Association for Computational Linguistics.

Klie, J. C., de Castilho, R. E., & **Gurevych, I.** (2020). From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains. *The 58th annual meeting of the Association for Computational Linguistics (ACL 2020)*, Virtual Conference. DOI: https://doi.org/10.18653/v1/2020.acl-main.624

Lander, Y., Maisak, T. A., & **Rakhilina, E. V.** (2012). Verbs of aquamotion: semantic domains and lexical systems. In M. Vulchanova & E. van der Zee E (Eds.), *Motion Encoding in Language and Space*. Oxford: Oxford University Press, 67–83. DOI: https://doi.org/10.1093/acprof:oso/9780199661213.003.0004

Leotta, R., & **De Felice, I.** (forth 2023). Metaphors and other figurative representations of the sea in Senecan works. In G. Volpe (Ed.), *Visuality: a crossroad between humanities and technologies*. Genoa: Genova University Press.

Lindenlauf, A. (2004). The sea as a place of no return in ancient Greece. *World Archaeology*, *35*(3), 416–33. DOI: https://doi.org/10.1080/0043824042000185801

Maisak, T. A., & **Rakhilina, E. V.** (2007). *Glagoly dvizhenija v vode: Leksicheskaja tipologija*. Moscow: Indrik.

Manning, P. (2015). World-Historical Gazetteer Research Report. *Journal of World-Historical Information*, *2–3*(1). DOI: https://doi.org/10.5195/jwhi.2015.21

Manning, P., & **Mostern, R.** (2015). World-Historical Gazetteer. DOI: https://doi.org/10.5195/jwhi.2015.21

Mostern, R. (2017). World-Historical Gazetteer Research Report. *Journal of World-Historical Information*, *4*(1). DOI: https://doi.org/10.5195/jwhi.2017.43

Nash, J. M. (2018). *Rulers of the Sea – Maritime Strategy and Sea Power in Ancient Greece 550–321 BC* (Doctoral dissertation, The Australian National University).

Nikoloska, A. (2012). The sea voyage of Magna Mater to Rome. *Histria Antiqua*, *21*, 365–71.

Reed, C. M. (2003). *Maritime traders in the ancient Greek world*. Cambridge, UK/New York: Cambridge University Pres.

Simon, R., Barker, E., & **Isaksen, L.** (2012). Exploring Pelagios: A Visual Browser for Geo-Tagged Datasets. *International Workshop on Supporting Users' Exploration of Digital Libraries*, Paphos (Cyprus), 1–6.

Simon, R., Isaksen, L., Barker, E., & **de Soto Cañamares, P.** (2016). The Pleiades Gazetteer and the Pelagios Project. In M. Lex Berman, R. Mostern, & H. Southall (eds.), *Placing Names: Enriching and Integrating Gazetteers*, 97–109. Bloomington: Indiana University Press. DOI: https://doi.org/10.2307/j.ctt2005zq7

Vitale, V., de Soto, P., Simon, R., Barker, E., Isaksen, L., & **Kahn, R.** (2021). Pelagios – Connecting Histories of Place. Part I: Methods and Tools. *International Journal of Humanities and Arts Computing, 15*(1–2), 5–32. DOI: https://doi.org/10.3366/ijhac.2021.0260

Wilkinson, T. (2020). *The Indian Ocean Trade and the Roman State* (MA dissertation, University of Wales Trinity Saint David).

TO CITE THIS ARTICLE:
Farina, A. (2023). Lost at Sea: A Dataset of 25+ SEA Words Morpho-Semantically Annotated in Ancient Greek and Latin. *Journal of Open Humanities Data*, 9: 24, pp. 1–7. DOI: https://doi.org/10.5334/johd.139

Submitted: 22 August 2023
Accepted: 26 October 2023
Published: 22 November 2023

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.