Journal of Open
Humanities Data

Translation Alignment for Ancient Greek: Annotation Guidelines and Gold Standards

COLLECTION:
REPRESENTING THE
ANCIENT WORLD
THROUGH DATA

DATA PAPER

]u[ubiquity press

CHIARA PALLADINO
FARNOOSH SHAMSIAN
TARIQ YOUSEF
DAVID J. WRIGHT
ANISE D'ORANGE FERREIRA
MICHEL FERREIRA DOS REIS

*Author affiliations can be found in the back matter of this article

ABSTRACT

This paper describes three datasets containing texts in Ancient Greek, manually aligned at word level against translations in English (Grc-Eng), Portuguese (Grc-Por) and Latin (Grc-Lat). The datasets were collected by two domain experts through annotation on the Ugarit Translation Alignment Editor (https://ugarit.ialigner.com/). The quality of each dataset was measured through Inter-Annotator Agreement (IAA) above 80%. Each dataset contains the aligned pairs and an Annotation Style Guide and serves as a Gold Standard for translation alignment of Ancient Greek, for the evaluation of automatic translation alignment models, and as high-quality training data. The Annotation Style Guide provides a starting point to approach the task of translation alignment for research and teaching. Data are stored on GitHub and Zenodo.

CORRESPONDING AUTHOR:
Chiara Palladino
Classics Department, Furman University, Greenville, South Carolina, US
chiara.palladino@furman.edu

KEYWORDS:
translation; computer applications; linguistic analysis

TO CITE THIS ARTICLE:
Palladino, C., Shamsian, F., Yousef, T., Wright, D. J., d'Orange Ferreira, A., & dos Reis, M. F. (2023). Translation Alignment for Ancient Greek: Annotation Guidelines and Gold Standards. *Journal of Open Humanities Data*, 9: 22, pp. 1–6. DOI: https://doi.org/10.5334/johd.131

(1) OVERVIEW

REPOSITORY LOCATION

GitHub: https://github.com/UgaritAlignment/Alignment-Gold-Standards.

Zenodo: Palladino, Shamsian & Yousef (2022b); Palladino, Wright & Yousef (2022); d'Orange Ferreira, Ferreira dos Reis & Yousef (2022).

Ugarit: The alignments can also be visualized on the Ugarit Website (https://ugarit.ialigner.com/).

CONTEXT

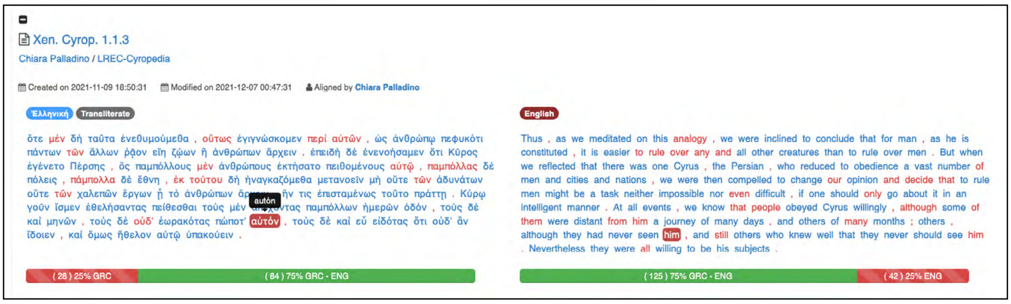
Data were produced as part of the research illustrated in Yousef et al. (2022a), Yousef et al. (2022b), and Yousef (2023). They are currently used in the following projects:

- For the evaluation of the performance of the Ugarit Automatic Alignment Model (https://huggingface.co/UGARIT/grc-alignment).
- To produce aligned datasets in the Ugarit Alignment Editor: https://ugarit.ialigner.com/.
- The Grc-Por Guidelines are being used to create new aligned corpora in the project "Letras clássicas digitais: interligando línguas antigas ao português e aprimorando um modelo automático de alinhamento de tradução" (Digital classics: linking ancient languages to Portuguese and enhancing an automatic model for translation alignment).

(2) METHODS

CREATING THE DATASETS AND THE GUIDELINES

Each dataset was created by two domain experts, who designed the Guidelines and annotated the corpus. The Grc-Eng Guidelines were created first and served as a model for the other two. Prior to aligning the corpus, the two domain experts created a first draft. Then, they aligned a subset to test the general consistency and feasibility of the Guidelines. For each new issue encountered in this phase there was a brief discussion, and a preferred annotation style was agreed upon. After the subset was completed, the experts completed the alignment without further discussion. The corpus was aligned using the Ugarit translation alignment editor (https://ugarit.ialigner.com/) (Figures 1–2).



Xen. Cyrop. 1.1.3
Chiara Palladino / LREC-Cyropedia

Created on 2021-11-09 18:50:31 Modified on 2021-12-07 00:47:31 Aligned by Chiara Palladino

Ἑλληνική Transliterate

ὅτε μὲν δὴ ταῦτα ἐνεθυμούμεθα , οὕτως ἐγιγνώσκομεν περὶ αὐτῶν , ὡς ἀνθρώπῳ πεφυκότι πάντων τῶν ἄλλων ῥᾷον εἴη ζῴων ἢ ἀνθρώπων ἄρχειν . ἐπειδὴ δὲ ἐνενοήσαμεν ὅτι Κῦρος ἐγένετο Πέρσης , ὃς παμπόλλους μὲν ἀνθρώπους ἐκτήσατο πειθομένους αὐτῷ , παμπόλλας δὲ πόλεις , πάμπολλα δὲ ἔθνη , ἐκ τούτου δὴ ἠναγκαζόμεθα μετανοεῖν μὴ οὔτε τῶν ἀδυνάτων οὔτε τῶν χαλεπῶν ἔργων ᾖ τὸ ἀνθρώπων ἄρ[illegible] ἤν τις ἐπισταμένως τοῦτο πράττῃ . Κύρῳ γοῦν ἴσμεν ἐθελήσαντας πείθεσθαι τοὺς μὲν [illegible]χοντας παμπόλλων ἡμερῶν ὁδόν , τοὺς δὲ καὶ μηνῶν , τοὺς δὲ οὐδ' ἑωρακότας πώποτ' αὐτόν , τοὺς δὲ καὶ εὖ εἰδότας ὅτι οὐδ' ἂν ἴδοιεν , καὶ ὅμως ἤθελον αὐτῷ ὑπακούειν .

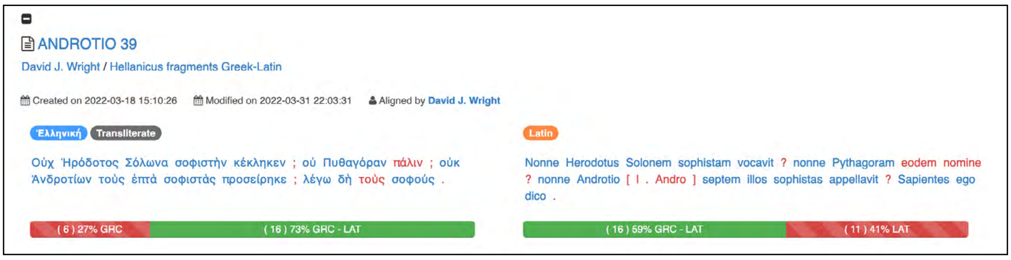
(28) 25% GRC (84) 75% GRC - ENG

English

Thus , as we meditated on this analogy , we were inclined to conclude that for man , as he is constituted , it is easier to rule over any and all other creatures than to rule over men . But when we reflected that there was one Cyrus , the Persian , who reduced to obedience a vast number of men and cities and nations , we were then compelled to change our opinion and decide that to rule men might be a task neither impossible nor even difficult , if one should only go about it in an intelligent manner . At all events , we know that people obeyed Cyrus willingly , although some of them were distant from him a journey of many days , and others of many months ; others , although they had never seen him , and still others who knew well that they never should see him . Nevertheless they were all willing to be his subjects .

(125) 75% GRC - ENG (42) 25% ENG

Figure 1 An example of a paragraph from Xenophon, *Cyropaedia*, aligned on Ugarit as part of the Gold Standard for Grc-Eng.



ANDROTIO 39
David J. Wright / Hellanicus fragments Greek-Latin

Created on 2022-03-18 15:10:26 Modified on 2022-03-31 22:03:31 Aligned by David J. Wright

Ἑλληνική Transliterate

Οὐχ Ἡρόδοτος Σόλωνα σοφιστὴν κέκληκεν ; οὐ Πυθαγόραν πάλιν ; οὐκ Ἀνδροτίων τοὺς ἑπτὰ σοφιστὰς προσείρηκε ; λέγω δὴ τοὺς σοφούς .

(6) 27% GRC (16) 73% GRC - LAT

Latin

Nonne Herodotus Solonem sophistam vocavit ? nonne Pythagoram eodem nomine ? nonne Androtio [I . Andro] septem illos sophistas appellavit ? Sapientes ego dico .

(16) 59% GRC - LAT (11) 41% LAT

Figure 2 An example of a fragment from the Digital Fragmenta Historicorum Graecorum, aligned on Ugarit as part of the Gold Standard for Grc-Lat.

The Grc-Por and Grc-Eng datasets consist of 2,010 words from the *Iliad*, 1,829 words from Plato's *Crito*, and 1,520 words from Xenophon's *Cyropaedia*, with the corresponding translations (Miller, 1914; Murray, 1924; Fowler, 1966; Campos, 2008; Cerdas, 2011; Werner, 2018). The corpus for Grc-Lat includes 100 fragments from the Digital Fragmenta Historicorum Graecorum

Project (Berti, 2021; 2023), with the corresponding Latin translation by philologist Karl Müller (Müller et al., 1841).

QUALITY CONTROL

To test the reliability and consistency of the resulting Guidelines and Gold Standard, we measured IAA over each dataset. IAA is considered when both annotators align the same pair of tokens or when both annotators do not align a token. Multi-word alignments (1–N, N–1, N–N) are flattened as 1–1 pairs. Let A_1 and A_2 be the flattened translation pairs created by each annotator, and I the intersection between them, we calculate IAA as follows:

$$IAA = 2 * I / (A_1 + A_2)$$

The resulting IAA was measured at 90.5% for Grc-Lat, 86.08% for Grc-Eng, and 83.31% for Grc-Por.

(3) DATASET DESCRIPTION

OBJECT NAME

Folders:

grc-eng

grc-lat

grc-por

Each folder contains:

- alignment_source_target.txt
- target.txt
- source-target-goldstandards.json
- source.txt
- guidelines_ source-target.pdf
- text_source_target.txt

FORMAT NAMES AND VERSIONS

We exported the Gold Standards in NAACL format (Mihalcea & Pedersen, 2003). This format allows the retrieval of translation pairs and corresponding sentences, but it also makes the corpus of parallel sentences available, in addition to the word-level alignments. Each dataset consists of the following files:

- Source sentences: source.txt (e.g. grc.txt).
- Target sentences: target.txt (e.g. eng.txt).
- The file source-target-goldstandards.json (e.g. grc-eng-goldstandards.json) contains the Gold Standard in JSON format. It provides the complete aligned sentences and translation pairs. Each entry in the JSON file includes two aligned source and target sentences and the complete sequence of translation pairs with unique IDs and link types, as they appear in the translation pairs file.
- The file alignment_source_target.txt (e.g. alignment_grc_eng.txt) contains the Gold Standard as a list of translation pairs. Each line in the file corresponds to a parallel sentence in the corpus. Each translation pair is identified through a source token ID within the source sentence, and a target token ID within the target sentence. Each translation pair is given a link type, *S* for Sure and *P* for Possible.
- The file text_source_target.txt (e.g. text_grc_eng.txt), contains the parallel sentences used in the Gold Standard, one pair per line, concatenated with the symbol |||.
- Alignment Guidelines are available in pdf, in English and Portuguese, in the format guidelines_source-target.pdf.

CREATION DATES

Start: 2022-01-19

End: 2022-11-14 (first release)

DATASET CREATORS AND CONTRIBUTIONS

1. Chiara Palladino: Grc-Eng, Grc-Lat Guidelines, dataset creation, conceptualization, data curation, writing – original draft, writing – revision.
2. Farnoosh Shamsian: Grc-Eng Guidelines, dataset creation, conceptualization, data curation.
3. Tariq Yousef: Gold Standard and IAA Calculation, formal analysis, software.
4. David J. Wright: Grc-Lat Guidelines, dataset creation, data curation.
5. Anise d'Orange Ferreira: Grc-Por Guidelines, dataset creation, data curation.
6. Michel Ferreira dos Reis: Grc-Por Guidelines, dataset creation, data curation.

LANGUAGES

Ancient Greek; Latin; English; Portuguese

LICENSE

Creative Commons Attribution 4.0 International https://creativecommons.org/licenses/by/4.0/legalcode

REPOSITORY NAME

GitHub: https://github.com/UgaritAlignment/Alignment-Gold-Standards

PUBLICATION DATE

2022-11-14

(4) REUSE POTENTIAL

Word-level manual alignments are extremely rare and challenging to create. Moreover, while there are some Alignment Guidelines available for modern languages (see https://ugarit.ialigner.com/guidelines.php for a partial list), these are currently the first ones explicitly addressing an ancient corpus, which has a set of specific problems usually not considered in modern languages. They provide an important contribution to NLP, but most importantly they define a workflow and the conceptual foundations to develop more aligned corpora of ancient and historical languages with their translations. Below are some suggested applications to reuse these datasets.

The Gold Standards provide a reliable, high-quality dataset to test and train automatic translation alignment models. These datasets are essential to evaluate the performance of such models and can be used as references to compare their predictions to assess their quality and reliability. We used them to evaluate the first transformer-based Translation Alignment model for ancient languages (Yousef et al., 2022a; Yousef et al., 2022b), which is a multilingual model, but they can also be used to evaluate the performance of monolingual models focusing on Ancient Greek. For example, the Grc-Por dataset and guidelines are currently being reused in the context of the project "Letras clássicas digitais: interligando línguas antigas ao Português e aprimorando um modelo automático de alinhamento de tradução", to improve the automatic alignment of Ancient Greek to Portuguese.

The aligned corpora can also be used to train AI models, both multilingual and monolingual, to perform other tasks, such as word sense disambiguation, Named Entity Recognition, and annotation projection. Yousef et al. (2023) and Yousef, Palladino & Jänicke (2023) offer insights on how to use these datasets for similar tasks.

The Guidelines provide a conceptual reference for people who wish to design or create aligned corpora in ancient languages. Being the first ones that explicitly address ancient texts, they cover issues such as controversial translations, fragmentary evidence, uncertainty, and phenomena connected to inflection. As they are not project-specific, they can be expanded and adapted depending on context. For example, they can be reused by scholars who use translation alignment for research, to provide an out-of-the-box reference to create a consistent corpus (Palladino, Shamsian & Yousef, 2022a). The Guidelines can also be adapted by teachers who use

Ugarit in the classroom to create aligned corpora for tests and assignments (Palladino, 2020; Palladino, Foradi & Yousef 2021; Shamsian & Crane, 2022). They can provide a general reference on how to handle typical linguistic phenomena that tend to be translated inconsistently, such as the genitive absolute, the use of the dative, proverbial expressions, or changes in verbal tense and voice. Using these guidelines, students can also be instructed to create consistent lexicographic indexes generated from the alignment of ancient texts.

Fundamentally, the Guidelines serve as a conceptual reference to create new sets of guidelines for other languages or contexts. Students and teachers can use analogous principles and select relevant phenomena, to create a more specific style guide that fits their needs. Moreover, scholars may use the general criteria and strategies illustrated in the guidelines to design new ones. This is currently being done in the project "Creating a corpus of Akkadian inscriptions with Ugarit", which aims at the creation of an aligned corpus of Akkadian and English, and in the Beyond Translation project for the creation of a Latin-English aligned text of the *Bellum Alexandrinum* (Crane et al., 2023).

ACKNOWLEDGEMENTS

We are grateful to Monica Berti and Gregory Crane for supporting the project, and to all the users of Ugarit.

FUNDING INFORMATION

The Ugarit project was generously supported by the Humboldt Chair of Digital Humanities at the University of Leipzig for the years 2016–2017.

The Grc-Por guidelines are currently being used to create new aligned corpora in the project "Letras clássicas digitais: interligando línguas antigas ao Português e aprimorando um modelo automático de alinhamento de tradução", funded by a FAPESP/e-Science grant (proc. n. 22/09490-0), for the years 2023–2025.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Chiara Palladino orcid.org/0000-0002-1811-5602
Classics Department, Furman University, Greenville, South Carolina, US

Farnoosh Shamsian orcid.org/0000-0003-3743-4278
Department of History, University of Leipzig, Leipzig, DE

Tariq Yousef orcid.org/0000-0001-6136-3970
Department of Mathematics and Computer Science, University of Southern Denmark, Odense, DK

David J. Wright orcid.org/0009-0007-0858-7458
Classics Department, Bowdoin College, Brunswick, Maine, US

Anise d'Orange Ferreira orcid.org/0000-0001-5755-1434
Faculdade de Ciências e Letras, Universidade Estadual Paulista (UNESP), Araraquara, São Paulo, BR

Michel Ferreira dos Reis orcid.org/0000-0003-2018-4188
Linguística e Língua Portuguesa, Universidade do Estado de Mato Grosso (UNEMAT), Pontes e Lacerda, Mato Grosso, BR

REFERENCES

Berti, M. (2021). *Digital Editions of Historical Fragmentary Texts*. Heidelberg: Propylaeum. DOI: https://doi.org/10.11588/PROPYLAEUM.898

Berti, M. (2023). Digital Fragmenta Historicorum Graecorum Project (DFHG). Retrieved August 23, 2023 from https://www.dfhg-project.org/.

Campos, A. M. (2008). *Apologia de Sócrates: precedido de Sobre a piedade (Êutifron) e seguido de Sobre o dever (Críton)*. Porto Alegre: L&PM. https://repositorio.usp.br/item/001674823.

Cerdas, E. (2011). *A Ciropedia de Xenofonte: um romance de formação na antiguidade*. São Paulo: Cultura Acadêmica. https://repositorio.unesp.br/handle/11449/109163.

Crane, G., Babeu, A., Cerrato, L. M., Parrish, A., Penagos, C., Shamsian, F., Tauber, J., & **Wagner, J.** (2023). Beyond Translation: Engaging with Foreign Languages in a Digital Library. *International Journal on Digital Libraries*, *24*, 163–176. DOI: https://doi.org/10.1007/s00799-023-00349-2

d'Orange Ferreira, A., Ferreira dos Reis, M., & **Yousef, T.** (2022). Critérios ou Convenções de Alinhamento do Grego às Traduções em Português (1.0) [Data set]. Zenodo. DOI: https://doi.org/10.5281/zenodo.7981097

Fowler, H. N. (1966). *Plato in Seven Volumes* (Vol. 1). Cambridge/London: Harvard University Press, William Heinemann Ltd.

Mihalcea, R., & **Pedersen, T.** (2003). An Evaluation Exercise for Word Alignment. *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, 1–10. https://aclanthology.org/W03-0301. DOI: https://doi.org/10.3115/1118905.1118906

Miller, W. (1914). *Xenophon in Seven Volumes* (Vol. 5–6). Cambridge/London: Harvard University Press, William Heinemann Ltd.

Müller, K., Müller, T., & **Langlois, V.** (1841). *Fragmenta Historicorum Graecorum* (5 vols.). Paris: Ambroise Firmin-Didot.

Murray, A. T. (1924). *The Iliad*. Cambridge/London: Harvard University Press, William Heinemann Ltd.

Palladino, C. (2020). Reading Texts in Digital Environments: Applications of Translation Alignment for Classical Language Learning. *The Journal of Interactive Technology and Pedagogy*, *18*. https://jitp.commons.gc.cuny.edu/reading-texts-in-digital-environments-applications-of-translation-alignment-for-classical-language-learning/.

Palladino, C., Foradi, M., & **Yousef, T.** (2021). Translation Alignment for Historical Language Learning: A Case Study. *Digital Humanities Quarterly*, *15*(3). http://www.digitalhumanities.org/dhq/vol/15/3/000563/000563.html.

Palladino, C., Shamsian, F., & **Yousef, T.** (2022a). Using Parallel Corpora to Evaluate Translations of Ancient Greek Literary Texts: An Application of Text Alignment for Digital Philology Research. *Journal of Computational Literary Studies*, *1*(1). DOI: https://doi.org/10.48694/jcls.100

Palladino, C., Shamsian, F., & **Yousef, T.** (2022b). Translation Alignment: Ancient Greek to English. Annotation Style Guide and Gold Standard (1.0) [Data set]. Zenodo. DOI: https://doi.org/10.5281/zenodo.7362097

Palladino, C., Wright, D. J., & **Yousef, T.** (2022). Translation Alignment: Ancient Greek to Latin. Annotation Style Guide and Gold Standard (1.0) [Data set]. Zenodo. DOI: https://doi.org/10.5281/zenodo.7981085

Shamsian, F., & **Crane, G.** (2022). Open Resources for Corpus-Based Learning of Ancient Greek in Persian. *The Journal of Interactive Technology and Pedagogy*, *21*. https://cuny.manifoldapp.org/read/open-resources-for-corpus-based-learning-of-ancient-greek-in-persian/section/2a96a89e-a4f8-4d47-bf78-f1d9a5b671bd.

Werner, C. (2018). *Homero, Ilíada* (1st ed.). São Paulo: Ubu Editora.

Yousef, T. (2023). *Translation Alignment Applied to Historical Languages* (PhD Thesis, University of Leipzig, Leipzig). DOI: https://doi.org/10.13140/RG.2.2.15623.57764

Yousef, T., Palladino, C., Heyer, G., & **Jänicke, S.** (2023). Named Entity Annotation Projection Applied to Classical Languages. *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 175–182. https://aclanthology.org/2023.latechclfl-1.19. DOI: https://doi.org/10.18653/v1/2023.latechclfl-1.19

Yousef, T., Palladino, C., & **Jänicke, S.** (2023). Transformer-Based Named Entity Recognition for Ancient Greek. *Digital Humanities 2023. Book of Abstracts*, 420–422. DOI: https://doi.org/10.5281/zenodo.8210808

Yousef, T., Palladino, C., Shamsian, F., d'Orange Ferreira, A., & **Ferreira dos Reis, M.** (2022a). An Automatic Model and Gold Standard for Translation Alignment of Ancient Greek. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5894–5905. https://aclanthology.org/2022.lrec-1.634.

Yousef, T., Palladino, C., Wright, D. J., & **Berti, M.** (2022b). Automatic Translation Alignment for Ancient Greek and Latin. *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, 101–107. https://aclanthology.org/2022.lt4hala-1.14.

TO CITE THIS ARTICLE:
Palladino, C., Shamsian, F., Yousef, T., Wright, D. J., d'Orange Ferreira, A., & dos Reis, M. F. (2023). Translation Alignment for Ancient Greek: Annotation Guidelines and Gold Standards. *Journal of Open Humanities Data*, 9: 22, pp. 1–6. DOI: https://doi.org/10.5334/johd.131

Submitted: 17 August 2023
Accepted: 17 October 2023
Published: 10 November 2023