GREEK, LATIN AND AUGMENTED INTELLIGENCE: THE OTHER AI

INTRODUCTION

This Profile looks at two technologies that were developed to make source texts in the original Greek, Latin and, indeed, any language directly accessible to audiences who have not yet studied – and may never study – the language itself: (1) translations aligned at the word and phrase level with the original text and (2) rich linguistic annotations explaining the part of speech, regularised dictionary form and syntactic function of each word in a corpus (typically called treebanks, because the syntactic structure is commonly visualised as an inverted tree).

A generation of students have grown up using reading environments that link inflected Greek and Latin forms (e.g. *fecit*) to morphological analyses (e.g. '*fecit*, 3rd person singular perfect indicative of *faciō*') and then to dictionary entries (e.g. short definitions as well as full entries from openly licensed lexica of Greek and Latin). The Perseus Digital Library (http://www.perseus.tufts.edu/hopper/), the interlinked Perseus under Philologic (https://perseus.uchicago.edu/) and Logeion systems at the University of Chicago (https://logeion.uchicago.edu/), and the Alpheios Reading Tools (https://alpheios.net/) provide these services for openly licensed content while the *Thesaurus Linguae Graecae* (https://stephanus.tlg.uci.edu/) provides more extensive coverage for Greek, but places most of its content behind a paywall. Similarly, the *Thesaurus Linguae Latinae* (https//thesaurus.badw.de/en/tll-digital/tll-open-access.html) provides an open access searchable PDF version (currently of volumes A–M, O–P and parts of N and R) but the full online database must be purchased through De Gruyter.

Translation alignments and treebanks are machine-actionable publications that have fundamentally new properties only possible in a digital format such as computability and extensibility. Both, however, have deep historical roots in print culture, and both can be seen as logical extensions of earlier formats. First, unlike printed interlinear translations, digital word and phrase level alignments can be analysed automatically, enhancing our ability to study the relationship between source text and translation. These relationships include changes in syntax (e.g. a translation of a Greek dative of possession such as 'there is to me a child' into an active construction such as 'I have a child'), translation equivalents for a given source text (e.g. the Greek word *mēnis* translated as 'anger' in one context vs 'wrath' in another) and source text varieties for a given translation (e.g. the English word 'anger' may translate not only *mēnis* but also *cholos*). Thus, the opening two words of the *Iliad*, *mēnin aeide*, can be translated as 'sing[*aeide*] the[0] divine wrath[*mēnin*]', with the article clearly marked as an addition with no corresponding word in the Greek. Alignments can link individual words or phrases (e.g. 'divine wrath' aligns to *mēnin*).

Second, a print commentary can explain the part of speech and syntactic role of any word in a sentence. Treebanks offer annotations providing a standardised dictionary form, part of speech, syntactic role and dependency for each word in a corpus. Thus, a treebank will indicate that *mēnin* is the feminine accusative singular of the dictionary entry *mēnis* and that *mēnin* is the object and depends upon the second person singular imperative verb, *aeide*, 'sing'. We can then find all nouns labelled as objects of *aeidō*

The Classical Review (2025) 75.1 7–22 © The Author(s), 2025. Published by Cambridge University Press on behalf of The Classical Association

https://doi.org/10.1017/S0009840X25000356 Published online by Cambridge University Press

('I sing'), all verbs of which *mēnis* is the object. The morpho-syntactic data makes it possible to search for many fairly complex syntactic features (e.g. different kinds of conditionals). Adding new features can make treebanks more useful. We can, for example, identify people, places, groups etc. to which pronouns refer (i.e. 'his wrath', 'his' designating Achilles). For languages such as Greek and Latin, which do not require nominative pronouns for subjects, we will also need to identify the subjects of verbs. This class of annotation (called coreference resolution) allows us to compare the semantics of particular named entities (e.g. all verbs of which Achilles and Hector are subjects in the *Iliad*, all adjectives associated with Greeks and Persians in Herodotus).

We can use manually created translation alignments and treebanks as training data for machine learning. We can then automatically align and linguistically analyse arbitrarily large corpora. The automatically generated alignments and linguistic annotations contain errors, but, once those errors are analysed, the automatic data can be used for larger-scale study. Likewise, human annotators can use automatically generated data as an initial dataset, which they can correct, with the corrected annotations available as new training data.

Work on translation alignments and on treebanks for Greek and Latin has been underway since 2007 and 2002 respectively,[1] but these two forms of publication, always intended to be combined so that they could easily reinforce each other, had been available only in separate environments. This Profile illustrates how translation alignments and treebanks can be integrated into a single reading environment.

Translation alignments and treebanks both open up new opportunities for research[2] (e.g. in translation studies, linguistics and more general philological analysis), but they are most significant because they help address the fact that very few readers will study languages such as Greek or Latin, much less maintain a high level of proficiency over time. Translation alignment and treebanks build upon specific applications of machine

[1]For a description of the need for treebanks, see G. Crane, 'Don't miss the lexicographers for the treebanks Philology in an electronic age', *Conference on the Cambridge Greek Lexicon* (2002), https://www.academia.edu/82054421/Dont_miss_the_lexicographers_for_the_treebanks_Philology_in_an_electronic_age. Five years later, in 2007, D. Bamman would begin working on alignments of Greek and Latin with English translations. The following year, 2008, H. Diakoff of the Alpheios Project recognised the value of manually aligning source texts and translations. Diakoff led development of a manual translation alignment editor.

[2]For translation alignment see C. Palladino and T. Yousef, 'To say almost the same thing? A study on cross-linguistic variation in ancient texts and their translations', *Digital Scholarship in the Humanities* 38 (2023), 1200–13, https://doi.org/10.1093/llc/fqac086; M. Alharbi et al., 'TransVis: Integrated Distant and Close Reading of Othello Translations', *IEEE Transactions on Visualization and Computer Graphics* 28 (2022), 1397–414, https://doi.org/10.1109/TVCG.2020.3012778; A. Fraisse, 'A Multilingual Dashboard to Analyse Intercultural Knowledge Circulation', in: O. Alonso et al. (edd.), *Linking Theory and Practice of Digital Libraries* (2023), pp. 8–14, https://doi.org/10.1007/978-3-031-43849-3_2; and for recent work with treebanks see J. Kostkan et al., 'OdyCy – A general-purpose NLP pipeline for Ancient Greek', in: S. Degaetano-Ortlieb et al. (edd.), *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (2023), pp. 128–34, https://aclanthology.org/2023.latechclfl-1.14; F. Gamba and D. Zeman, 'Universalising Latin Universal Dependencies: A Harmonisation of Latin Treebanks in UD', in: L. Grobol and F. Tyers (edd.), *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)* (2023), pp. 7–16, https://aclanthology.org/2023.udw-1.2.

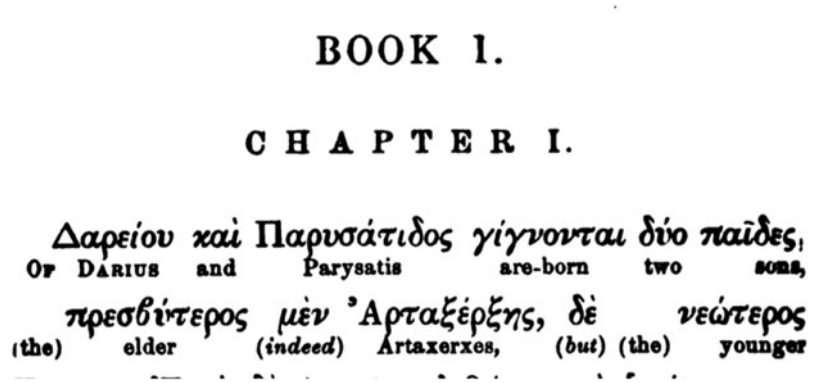
https://doi.org/10.1017/S0009840X25000356 Published online by Cambridge University Press

learning[3] (e.g. systems trained to align and parse) and of emergent artificial intelligence (systems that learn how to align and parse without specifically being trained to do so). But translation alignment and treebanks are primarily important because they are examples of the 'other AI': augmented intelligence.[4] This second AI can open up specialist knowledge about subjects such as Greek and Latin to new audiences and transform the role that a field such as Graeco-Roman culture can play in the intellectual life of humanity.

The rest of this Profile provides a brief history of translation alignment and treebanks as separate activities, and then discusses work done to bring translation alignment and treebanks together in an integrated reading environment.

TRANSLATION ALIGNMENT

Aligning words and phrases in a translation with its source text has a long tradition. Conventional translations can include untranslated terms and English approximations side-by-side to emphasise that the source term cannot be properly translated (e.g. the opening of the *Iliad*: 'Anger [*mēnis*], goddess, sing it, of Achilles, son of Peleus— 2 disastrous [*oulomenē*] anger that made countless pains [*algea*] for the Achaeans')[5]. Interlinear translations for Greek and Latin texts appeared in the nineteenth century and mainly served readers who were forced to study these languages to fulfil various academic requirements and wanted to learn as little grammar as possible (e.g. fig. 1).


BOOK I.

CHAPTER I.

Δαρείου καὶ Παρυσάτιδος γίγνονται δύο παῖδες,
OF DARIUS and Parysatis are-born two sons,
πρεσβύτερος μὲν Ἀρταξέρξης, δὲ νεώτερος
(the) elder (indeed) Artaxerxes, (but) (the) younger

FIGURE 1. Opening of Xenophon's *Anabasis*, in: T. Clark, *The Anabasis of Xenophon: with an interlinear translation* [1859], p. 9 (https://hdl.handle.net/2027/uc2.ark:/13960/t1sf2p143?urlappend=%3Bseq=15).

[3]For one example of the use of machine learning in classical philology see B. Graziosi et al., 'Machine Learning and the Future of Philology: A Case Study', *TAPA* 153 (2023), 253–84, https://doi.org/10.1353/apa.2023.a901022.

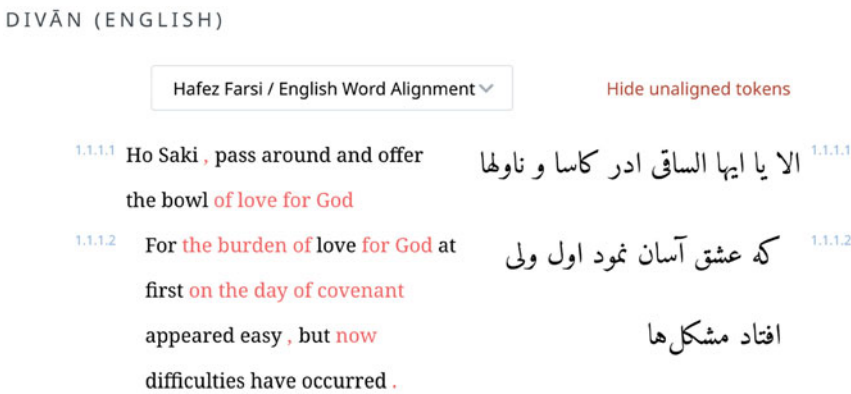
[4]For more on a definition of augmented intelligence (first defined by D.C. Engelbart in 1962) see D.C. Engelbart, 'Augmenting Human Intellect: A Conceptual Framework', in D. Araya and P. Marber (edd.), *Augmented Education in the Global Age* (2023), pp. 13–29, https://doi.org/10.4324/9781003230762; M. Pasquinelli, 'Augmented Intelligence', *Critical Keywords for the Digital Humanities* (2014), https://matteopasquinelli.com/augmented-intelligence/; M.N.O. Sadiku and S.M. Musa, 'Augmented Intelligence', in: M.N.O. Sadiku and S.M. Musa, *A Primer on Multiple Intelligences* (2021), pp. 191–9.

[5]The Center for Hellenic Studies Homeric *Iliad*: https://chs.harvard.edu/primary-source/homeric-iliad-sb/.

https://doi.org/10.1017/S0009840X25000356 Published online by Cambridge University Press

Contemporary translation alignment has developed along three complementary pathways. First, automatic methods generate word and phrase level alignments for source texts and translations that have been aligned at a coarse level, such as a page, paragraph or sentence. Although automatic alignment at the word and phrase level has a substantial error rate, the ability to process large bodies of textual data means that meaningful signals emerge from the noise. Automatic alignment is a fundamental tool for analysing translations at scale.

Second, manual alignments between source texts and translations can shed surprising light on the relationship between the two. The figure below provides an example from Persian poetry, a language with which many (and probably most) readers of this piece are not familiar. The underlined English words in figure 2 do not have equivalents in the Persian source.



1.1.1	الا یا ایها الساقی ادر کاسا و ناولها
	Ho Saki, pass around and offer the bowl of love for God
1.1.2	که عشق آسان نمود اول ولی افتاد مشکل‌ها
	For the burden of love for God at first on the day of covenant appeared easy, but now difficulties have occurred.

FIGURE 2. Underlined words do not correspond to anything in the Persian and add a dimension of Neoplatonic allegory.[6]

[6]The alignments were manually produced as a part of M. Foradi, 'Engagement with Classical Literature in the Framework of a Citizen Science Project Using Translation Alignment: Data Accuracy and Pedagogical Effectiveness' (Diss., Leipzig University, 2020). The figure was taken from the initial version of Beyond Translation, https://beyond-translation.perseus.org. The translation is from H.W. Clarke, *The Divan by Hafez: Translated for the First Time out of the Persian into English Prose* (1891), https://archive.org/details/thedivan01hafiuoft.

https://doi.org/10.1017/S0009840X25000356 Published online by Cambridge University Press

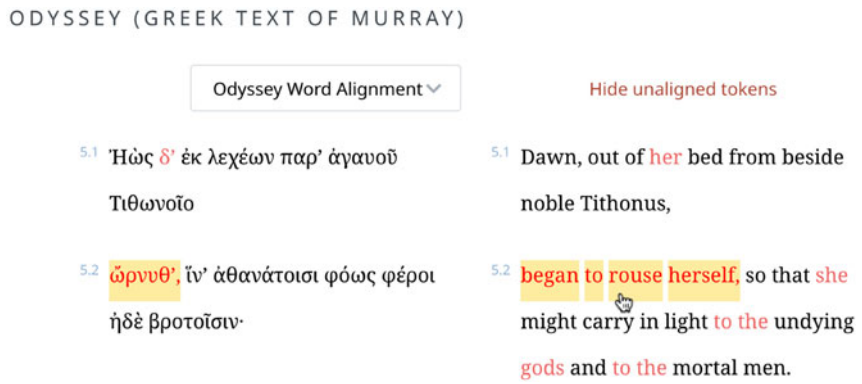


FIGURE 3. Born-digital translation of *Odyssey* 5.1–2 with unaligned words and selected alignment. Words that do not have equivalents in the other language are red. In the Greek, the only word that does not have an equivalent is δ᾽ in 5.1. The English has added ‘her’ in 5.1 and ‘she’, ‘to the’, ‘gods’ and ‘to the’ in line 5.2.

The translator, Henry Wilberforce Clarke, has freely, and without explanation, added extensive religious language to convert a poem about drinking and sexual desire into a Neoplatonic allegory. No one without an understanding of the Persian text could have seen those additions in a traditional printed text. Anyone with a knowledge of English, whether or not they know Persian, can see words in the aligned English translation that have no counterparts in the Persian.

Third, translations can be designed from the start to align with and shed light upon the original source text. Such translations can serve as intermediaries for those who wish to understand more deeply the relationship between a more or less free literary translation and the original source text. Such translations can also serve those who wish to learn the language and can intentionally make particular features more prominent.

Figure 3 shows a born-digital translation, developed by A. Parrish and G. Crane, for the opening of *Odyssey* 5. The unaligned words (shown in red) now more narrowly reflect linguistic features of Greek (e.g. Greek does not need to include possessive pronouns, and the definite article is far less commonly used in Homer than in later Greek). The selected alignment also shows an effort to illustrate two grammatical features of the verb *ōrnuth'* (from *ornumi*, ‘to rouse, set in motion’): the verb form is middle (hence ‘to rouse herself’) and an imperfect (‘began to’ suggests that this action takes place over a period of time, though admittedly brief). Those learning Greek should be able to scan through a text and see each middle and each imperfect form glossed in such a way that the linguistic features are prominent.

The earliest parallel text alignment work focused on automatically aligning words in source text and translation. Words that tended to occur in both source text and translation were often translation equivalents, and parallel text analysis was used to support machine translation systems of the time.[7] A decade later, D. Bamman, then a researcher at Perseus,

[7]For example, P.F. Brown et al., ‘The Mathematics of Statistical Machine Translation: Parameter Estimation’, *Computational Linguistics* 19 (1993), 263–311; F. Smadja et al., ‘Translating Collocations for Bilingual Lexicons: A Statistical Approach’, *Computational Linguistics* 22 (1996), 1–38.

https://doi.org/10.1017/S0009840X25000356 Published online by Cambridge University Press

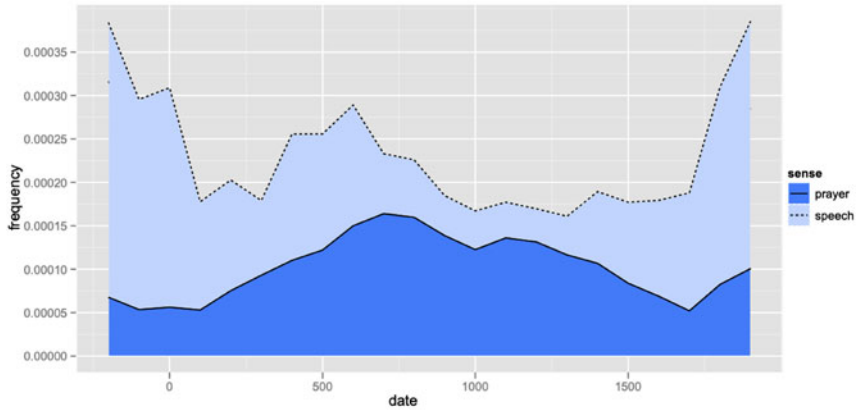


FIGURE 4. English sense distribution for the Latin word *oratio* (96,313 instances).[8]

applied this method to the parallel Greek/English and Latin/English corpora available in Perseus at the time. Distinguishing word senses is a notoriously difficult task, and one approach suggested that a good way to identify different word senses was to track places where translators chose very different translation terms. The Latin word *oratio* can be translated as 'speech', when it describes a secular oration, and as 'prayer', when it describes a religious act (see fig. 4). Bamman was able to trace the changing frequency of these word senses (and the changing importance of religious vs secular contexts) in Latin sources over more than a thousand years.

Automatic translation alignment only identifies a subset of translation pairs, and it has a substantial error rate (in Bamman and Crane [n. 8] 22% of the proposed translation equivalents for *oratio* were false); but errors tended to be random, and the overall patterns often emerged clearly from the background noise.

H. Diakoff of the Alpheios Project, however, recognised the potential value of manual alignment of source text and translation, both because the data would be more precise and because the exercise of aligning source text and translation seemed to be a pedagogically useful exercise for language learners, both by seeing how fuzzy the relationship could be between sources and translations and by creating their own translations for alignment with the source text (where learners could see, for example, if they had left words untranslated). The Alpheios Project began developing a manual translation alignment tool in 2009.[9] Over ten years later M. Foradi at the University of Leipzig studied translation alignment (Foradi, n. 6). Two of her findings were particularly suggestive. First, faculty (including one of the co-authors) had been asking students to align an English translation to a source text in an unknown language (Greek or Latin) by using basic scaffolding tools (e.g. the ability to click on any inflected word in a text and see a dictionary entry). She performed an

[8]See D. Bamman and G. Crane, 'Measuring historical word sense variation', *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (2011), pp. 1–10, https://doi.org/10.1145/1998076.1998078, Figure 5.

[9]Foradi (n. 6) discusses the earliest version of the Alpheios alignment editor at length. For the most recent version of the Alpheios Alignment editor see https://alignment.alpheios.net/whats-new.html; https://www.youtube.com/watch?v=AJwd_JXLL5Q.

https://doi.org/10.1017/S0009840X25000356 Published online by Cambridge University Press

experiment to see how well readers could align a translation to a source text in an unknown language. She provided word/phrase level alignments between Persian and pre-existing, rather free English translations to German students who did not know Persian (but who had studied Arabic and could easily work with the slightly different form of Arabic script used for Persian). These students created alignments to a German translation that were just as accurate as those made by experts in both languages (Persian speakers who were fluent in German). This case study reflects a general use case: a reader uses alignments between the source text and an openly licensed translation to scrutinise the relationship between a new literary translation and the source.

Second, and perhaps more surprisingly, Foradi wanted to see if students learned Persian vocabulary as a by-product of aligning Persian and German. She had a control group of students spend equal time studying vocabulary with flashcards. The flashcard group did better on a vocabulary test on the following day, but the alignment group did better two months later. The alignment exercise thus was less efficient in the short run, but had more lasting cognitive impact. Foradi's work laid the foundation for further research (see below).

In 2016 computer scientist T. Yousef, in collaboration with digital humanist C. Palladino, developed Ugarit, a web-based translation alignment system in which 797 users have created 1,215,828 manual word/phase level alignments in 33,595 texts.[10] Formal guidelines now exist for aligning ancient Greek with Latin, English with Portuguese, and Latin with English.[11] Ugarit benefited from a particularly well-designed user interface, and it also illustrates a crucial functionality for one class of ongoing work. Ugarit can display alignments between a source text and two translations, allowing editors to juxtapose an intentionally literal translation, designed to reveal as much as possible about the Greek source, with a literary translation. This three-text mode can spur a new class of review designed to make the translator's decisions accessible to a much wider audience.

The scale of the curated alignments has had two effects. First, we can study translation pairs with greater precision than was possible with the more extensive, but noisier, alignment data that Bamman was able to generate a decade earlier.

Second, Yousef had enough trusted data to train models for a new generation of automatic alignment between Greek and Latin, Greek and English, and Latin and English. This new generation of automatic alignment also benefits from access to language models that, in turn, can recognise words with similar meanings. Thus, an aligner using such a language model can infer that, if Greek *lithos* aligns with English 'rock', then it can align with English 'stone' as well. Yousef[12] reported that 91.5% of proposed alignments between Greek and English were correct and that the automatic alignments identified

[10]https://ugarit.ialigner.com/guidelines.ph.

[11]T. Yousef et al., 'An automatic model and Gold Standard for translation alignment of Ancient Greek', in: N. Calzolari et al. (edd.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (2022), pp. 5894–905, https://aclanthology.org/2022.lrec-1.634, and C. Palladino et al., 'Translation Alignment for Ancient Greek: Annotation Guidelines and Gold Standards', *Journal of Open Humanities Data* 9 (2023), https://doi.org/10.5334/johd.131.

[12]T. Yousef et al., 'Automatic Translation Alignment for Ancient Greek and Latin', in: R. Sprugnoli et al. (edd.), *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages* (2022), pp. 101–7, https://aclanthology.org/2022.lt4hala-1.14.

https://doi.org/10.1017/S0009840X25000356 Published online by Cambridge University Press

87.3% of all possible correct alignments.[13] Results will vary depending on how close the translation is to the source text and on how fine-grained the intermediate chunks are (e.g. 25-word sentences vs 75-word sections), but the larger picture is clear. We can now generate far more accurate alignments between Greek, Latin and English texts and perform more comprehensive semantic analysis than was possible before. We also can generate automatic alignments that are close enough to be helpful for readers who do not know Greek or Latin.

Using Ugarit poses at least one challenge: once a translation has been entered into Ugarit, it cannot be edited. If readers discover that there was an error in transcribing a pre-existing translation or if they wish to revise a translation of their own while aligning it, they have to start over. A. Parrish (Tufts '21) and Crane worked on a text-only alignment format, in which Greek words are interspersed with their translations.[14] The format is not as visually as attractive as that of Ugarit, but it provides great flexibility for ongoing revision.

Much of the most interesting recent work focuses on applications of translation alignment to language learning. C. Palladino documents the effects of having students align one or more existing translations to a source text, forcing them to focus on each translation much more precisely than would normally be the case. She also described the strategy of having students produce aligned translations of their own, a practice that makes it easier for them to see where they may have missed a word or a phrase and for their instructor to see the precise relationship between source text and student translation.[15] Large language models that can recognise synonymous expressions should also now be able to provide immediate feedback, identifying places where students produce semantically improbable translations. F. Shamsian has, as part of a Leipzig Ph.D. dissertation, developed a localisable infrastructure for learning ancient Greek. Her immediate learner language is Persian, but the larger goal is to create a compact, easily translated and localised framework that can support the study of ancient Greek in many languages (especially from outside of Europe). She reports on how, after 30 hours of instruction, Persian speakers were able to critique existing Persian translations of the *Iliad* and create more accurate representations of the Greek original. The same learners were, at this point, able to begin producing the first direct translations of Plato's *Crito* into Persian. As a result, we have three finalised translations of *Crito* under development that each represent a different interpretation or approach to the text.[16] Most recently, J. Hilleary studied readers using translation alignment to work with a source text in Estonian, a contemporary non-Indo-European language with which none of the participants were familiar. Half the participants had an English translation aligned at the sentence level, and the other half had access to word and phrase level alignments. All participants were

[13]In information retrieval these figures would be described as precision (91.5% of the retrieved items were correct) and recall (87.3% of the correct items were found).

[14]https://sites.tufts.edu/perseusupdates/2022/03/25/perseus-and-new-enhanced-introductions-to-ancient-greek-fall-2022/.

[15]C. Palladino, 'Reading Texts in Digital Environments: Applications of Translation Alignment for Classical Language Learning', *The Journal of Interactive Technology and Pedagogy* 18 (2020), https://cuny.manifoldapp.org/read/reading-texts-in-digital-environments-applications-of-translation-alignment-for-classical-language-learning-310daf57-ed21-4150-b5ba-7386399ef905/section/5ae515f9-4e47-4a1e-a0c6-eb3b5f0b1a03.

[16]F. Shamsian and G. Crane, 'Open Resources for Corpus-Based Learning of Ancient Greek in Persian', *Journal of Interactive Technology and Pedagogy* 21 (2021), https://cuny.manifoldapp.org/read/open-resources-for-corpus-based-learning-of-ancient-greek-in-persian/section/2a96a89e-a4f8-4d47-bf78-f1d9a5b671bd.

https://doi.org/10.1017/S0009840X25000356 Published online by Cambridge University Press

able to generate reasonable answers to fill-in-the-blank ('cloze') reading comprehension questions in Estonian. Word-level alignments, as predicted, greatly enhanced performance, and participants were often able not only to identify the appropriate Estonian word, but also to generate new morphological forms in Estonian.[17] When asked what additional information they would have found useful, many participants asked for a tool to understand the structure of sentences in the story, a category of data to which we turn next in our discussion of treebanks.

TREEBANKS

Treebanks provide information that supports general linguistic research, but that also helps readers make much fuller use of translation alignments. Treebanks are textual databases that contain multiple linguistic annotations for each word. Consider the opening words of the Homeric *Iliad*: *mēnin aeide*, 'sing the wrath'. The core features are regularised dictionary form (e.g. *mēnin* is a form of the noun *mēnis*, 'wrath'), part of speech (e.g. *mēnin* designates the feminine accusative singular), syntactic role (*mēnin* is an object) and dependency (*mēnin* is governed by *aeide,* 'sing!'). They derive their name from the fact that they are a form of databank and because the syntactic analyses are often visualised in graphical form as upside-down trees.

Corpus linguists developed treebanks so that they could extract precise information about linguistic usage (e.g. the frequency subject-verb-object vs subject-object-verb order or different case uses with a given verb). Treebanks allow researchers to explore ideas and quantify their results, moving away from vague terms such as 'common in poetry' or 'typical of late Greek prose'.[18]

[17]J. Hilleary, 'Facilitating Language Hacking with Digital Tools: A Study of Translation Alignment' (Tufts University, Msc in Computer Science, 2024).

[18]Examples of research using treebank data include Bamman et al., 'A Case Study in Treebank Collaboration and Comparison: Accusativus Cum Infinitivo and Subordination in Latin', *Prague Bulletin of Mathematical Linguistics* 90 (2008), 109–22; D. Bamman and G. Crane (n. 8); G.G.A. Celano and G. Crane, 'Semantic Role Annotation in the Ancient Greek Dependency Treebank', in: M. Dickinson et al. (edd.), *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)* (2015), pp. 26–34; G.G.A. Celano, 'An Automatic Morphological Annotation and Lemmatization for the IDP Papyri', in: N. Reggiani (ed.), *Digital Papyrology II* (2018), pp. 139–48, https://doi.org/10.1515/9783110547450-008; G.G.A. Celano, 'The Dependency Treebanks for Ancient Greek and Latin', in: M. Berti (ed.), *Digital Classical Philology* (2019), pp. 279–98, https://doi.org/10.1515/9783110599572-016; G.G.A. Celano, 'Lemmatization and morphological analysis for the Latin Dependency Treebank', *Studi e Saggi Linguistici* 58 (2020), 21–38; R. Gorman, 'Author Identification of Short Texts Using Dependency Treebanks without Vocabulary', *Digital Scholarship in the Humanities* 35 (2020), 812–25, https://doi.org/10.1093/llc/fqz070; R. Gorman, 'Universal Dependencies and Author Attribution of Short Texts with Syntax Alone', *Digital Humanities Quarterly* 16 (2022), http://www.digitalhumanities.org/dhq/vol/16/2/000606/000606.html; F. Mambrini and M. Passarotti, 'Non-Projectivity in the Ancient Greek Dependency Treebank', *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)* (2013), pp. 177–86, https://aclanthology.org/W13-3720; F. Mambrini and M. Passarotti, 'Subject-Verb Agreement with Coordinated Subjects in Ancient Greek: A Treebank-Based Study', *Journal of Greek Linguistics* 16 (2016), 87–116, https://doi.org/10.1163/15699846-01601003; F. Mambrini and M. Passarotti, 'Linked Open Treebanks. Interlinking

https://doi.org/10.1017/S0009840X25000356 Published online by Cambridge University Press

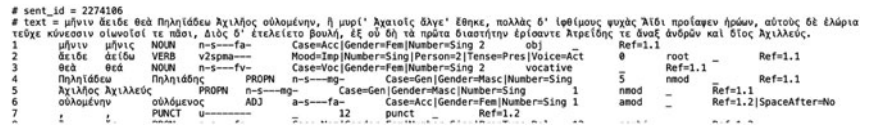


FIGURE 5. A treebank automatically produced by GLAUx in the Perseus Dependency Treebank, then edited and converted to the Universal Dependency Framework by the Daphne Treebank repository of Ancient Greek Poetry (https://perseids-publications.github.io/daphne-trees/).

Figure 5 illustrates the raw data behind a treebank created using the Perseus annotation guidelines and represented as XML.[19] The subject of the main verb *gignontai* (which is the root node of sentence) is *paides*, which is in turn modified by *duo* ('two children exist', 'there are two children'). The head attribute indicates dependency. The main verb is the root of the tree and has no dependency (thus, for *gignontai*: head = '0'). Other than the root, each word has one and only one head in the default case. (We can extend the syntax by creating additional links if we want to show that one word notionally depends on multiple words as in 'she saw and read many books'.) Any word can have multiple dependencies: thus, *duo*, 'two', and *Dareiou*, 'of Darius', both modify *paides*. The CO (for 'co-ordinated') relation or co-reference relation indicates that *Parysatis*, 'of Parysatis', is parallel to *Dareiou.* With data such as this we could answer questions such as which subjects or objects appear with any verb, how often a verb has no object, which adjectives modify which nouns, or how often subjects precede or follow the main verb.

Treebanks can, however, also show readers how each word fits into its sentence in the source text. Learners of ancient Greek and Latin have spent thousands of years answering questions about which word depends on which and why. Treebanks present much of that information in explicit form and allow readers to see the structure of sentences laid out in a diagram. Existing treebanks contain enough information to support a range of queries. Part of speech data allows readers to ask for particular forms (e.g. the distribution of tenses in histories vs drama, the relative frequency of future participles, a form typically used to express purpose, or the frequency of the subjunctive vs the optative over time). The dependencies allow more complex queries (e.g. the number of times that the Greek verb *echō*, 'to have', in a given author or genre has an object vs being intransitive and

Syntactically Annotated Corpora in the LiLa Knowledge Base of Linguistic Resources for Latin', in: M. Candito et al. (edd.), *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)* (2019), pp. 74–81, https://aclanthology.org/W19-7808/; F. Mambrini, 'Treebanking in the World of Thucydides. Linguistic Annotation for the Hellespont Project', *Digital Humanities Quarterly* 10 (2016), http://www.digitalhumanities.org/dhq/vol/10/2/000251/000251.html; F. Mambrini, 'Nominal vs copular clauses in a diachronic corpus of Ancient Greek historians', *Journal of Greek Linguistics* 19 (2019), 90–113; T. Van Hal and A. Keersmaekers, 'Visualizing the Ancient Greek Forest through the Trees: How Treebanks Can Advance the Education of Classical Languages', *Les Études Classiques* 89 (2021), 349–72; M. Vierros and E. Henriksson, 'PapyGreek Treebanks: A Dataset of Linguistically Annotated Greek Documentary Papyri', *Journal of Open Humanities Data* 7 (2021), 26, https://doi.org/10.5334/johd.55.

[19] http://perseusdl.github.io/treebank_data/.

https://doi.org/10.1017/S0009840X25000356 Published online by Cambridge University Press



FIGURE 6. A treebank for the opening of Xenophon's *Anabasis*.

designating a state or complex queries that can retrieve particular classes of conditionals by looking at the moods and tenses of verbs in primary and subordinate clauses). Additional categories of annotation can be added to support a wider range of queries (e.g. distinguishing between nouns that designate animate and inanimate entities or linking each verb in a document with its subject).

Figure 6 shows the automatically annotated opening sentence of Xenophon's *Anabasis*, produced by the GLAUx Trees and visualised in *Beyond Translation*. The arguments attached to the base URL (https://beyond-translation.perseus.org/reader/) are:

```
urn:cts:greekLit:tlg0032.tlg006.perseus-grc2:1.1.1?
mode=syntax-trees
```

The identifiers 'tlg0032' and 'tlg006' identify Xenophon and the *Anabasis*, while 'perseus-grc2' and '1.1.1' specify the edition used and the traditional citation (book 1, chapter 1, section 1).

G. Crane first wanted to be able to capture syntactic data when, in 1984, he was composing Morpheus, a rule-based system that analyses inflected Greek and Latin forms. Syntax, however, does not lend itself to rule-based programming. H. Kučera and N. Francis had, however, built the prototype for what Crane needed two decades earlier, creating a one-million-word parsed corpus of English.[20] Treebanks became more

[20]https://en.wikipedia.org/wiki/Brown_Corpus.

https://doi.org/10.1017/S0009840X25000356 Published online by Cambridge University Press

prominent in the 1990s, and in 2002 Crane suggested that a dictionary such as the *Cambridge Greek Lexicon* should be built on a treebanked corpus.[21] When D. Bamman joined Perseus four years later, his first task was to begin work on a treebank of Latin, for which he received funding from the National Science Foundation.[22] Bamman found that Czech linguists, working with their own highly inflected language with relatively free word order, had developed methods that were much better suited to languages such as ancient Greek and Latin than the methods developed for languages such as English and French, with relatively thin morphology and rigid word order.

Two other groups proved to be developing treebanks for Greek and Latin. The Proiel Project in Norway created treebanks for multiple language versions of the New Testament for comparative study.[23] The Index Thomisticus[24] had begun treebanking Thomas Aquinas. D. Bamman and M. Passarotti of the Index Thomisticus worked hard to fashion a nearly identical annotation scheme for both projects based on the Prague Dependency Treebank.[25] Proiel chose a scheme that differed a bit more, but still remained largely in agreement with the other two. Data from all three projects (and their derivatives) has been successfully combined.

Support from the Alpheios Project made it possible to create treebanks for Homeric and Hesiodic epic. Individuals annotated each sentence separately, and a more senior editor resolved places where the two independent annotations diverged. Each sentence thus has three credits. When we moved on to Greek drama, we identified an individual specialist, F. Mambrini, who created the first treebanks (and a unique edition) for the seven surviving plays of Aeschylus. Funding from the Institute for Museum and Library Services, the Mellon Foundation and the Alexander von Humboldt Foundation in Germany paid for the development of a dedicated Treebank editor, Arethusa, in 2013–2014. The Mellon-funded Perseids Project assumed management of Arethusa,[26] and the amount of treebanked Greek and Latin expanded over the following years, with more than 1.2 million words of Greek available in the Perseus and Proiel formats.[27]

Two major changes have taken shape in the intervening fifteen years. First, treebanks were used from the earliest stage of work as training data for machine-learning algorithms, and larger bodies of Greek and Latin were automatically treebanked. However, as the amount of training data has expanded and as advances in algorithms and, especially, increases in computing power have made machine-learning systems more capable, the quality of automatically generated treebanks has increased. The GLAUx Project[28] at Leuven has published treebanks in the Perseus Dependency Treebank format with more

[21]Crane (n. 1).
[22]National Science Foundation (BCS-0616521).
[23]http://dev.syntacticus.org/proiel.html.
[24]https://www.corpusthomisticum.org/it/index.age.
[25]See D. Bamman et al. (n. 18).
[26]https://www.perseids.org/tools/arethusa/app/#/.
[27]For an overview of much of this work as well as some later work see G.G.A. Celano (n. 18 [2019]); for a discussion of the challenges involved in integrating data from different treebanks see B. Hwang, 'Experiments in Digital Philology', *Perseus Journal of Data Preservation and Sustainability* (2023), https://pdldatajournal.pubpub.org/pub/article2.
[28]A. Keersmaekers, 'The GLAUx corpus: methodological issues in designing a long-term, diverse, multi-layered corpus of Ancient Greek', in: N. Tahmasebi et al. (edd.), *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021* (2021), pp. 39–50, https://doi.org/10.18653/v1/2021.lchange-1.6; https://github.com/alekkeersmaekers/GLAUx.

https://doi.org/10.1017/S0009840X25000356 Published online by Cambridge University Press

than 19 million words, and that figure can be increased to include any texts that they choose to process. The task for annotators has shifted from creating treebanks by hand to reviewing and augmenting automatically generated data. In this environment some annotators move away from elaborate editing environments and work directly with annotations in a text-only format (greatly simplifying the needed infrastructure).

Second, a newer annotation scheme, the Universal Dependency (UD) Framework[29], has emerged and achieved much broader adoption. UD was designed from the start for cross-language comparison, and it has developed a substantial community: more than 200 treebanks covering more than 100 languages are available in the UD framework – including 200,000 tokens of Greek converted by Perseus and Proiel each (more than 400,000 in total) into UD. Where those who became familiar with the Perseus Dependency Treebank annotation scheme could easily work with Czech, those who become comfortable with UD can move across more than 100 languages. Put another way, working closely with UD Greek and Latin treebanks provides readers with skills that are far more immediately and widely applicable than mastery of traditional Greek and Latin grammar.

At the same time published UD treebanks attracted the attention of the latest natural language processing pipelines: Stanza and SpaCy.[30] Because Perseus and Proiel published UD versions for parts of their treebanks, Stanza and SpaCy can automatically generate initial treebanks for Greek and Latin texts. Automatic parsing for Greek (based on just over 400,000 tokens) lags behind the automated parsing based on data in the Perseus/Prague scheme (based on 1.2 million tokens), but it will improve as new data becomes available in UD. Although treebanks for classical Latin lag behind those available for classical Greek, more than 1 million words of Latin from various periods are available in UD format, and automatic treebanking of Latin has begun to build upon this foundation.[31]

INTEGRATION OF TRANSLATION ALIGNMENTS AND TREEBANKS

Treebanks and translation alignments were designed to complement each other. D. Bamman used both resources to model a dynamic lexicon that could, more than fifteen years ago, automatically generate treebank and alignment data to provide basic lexicographic information for arbitrarily large corpora.

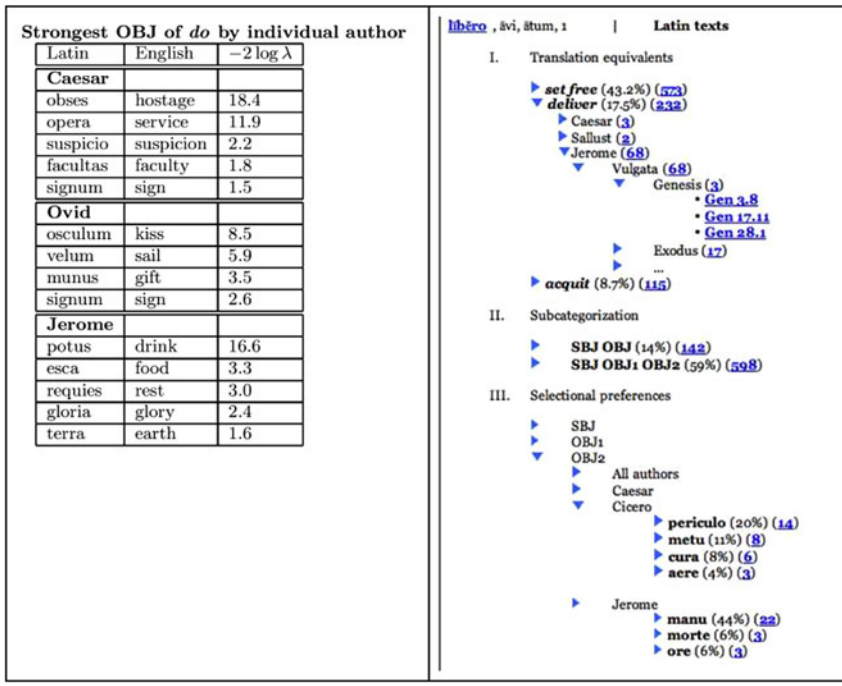
Figure 7 illustrates an early application that draws upon both treebanks and alignments. The table on the left first uses treebank data to identify objects of the verb *faciō*, 'I do', and then draws upon translation alignments to gloss the Latin words. The figure on the right provides a model for a dynamic lexicon entry that uses automatically generated translation alignments and treebanking. The methods for automatic alignment and treebanking were much less accurate at the time when this work was done than they are now, but the most common patterns emerged against the noisy background. We are now at a point,

[29]R. McDonald et al., 'Universal Dependency Annotation for Multilingual Parsing', *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (2013), pp. 92–7, https://aclanthology.org/P13-2017; https://universaldependencies.org/.

[30]https://stanfordnlp.github.io/stanza/; https://spacy.io.

[31]See P.J. Burns, 'LatinCy: Synthetic Trained Pipelines for Latin NLP', arXiv:2305.04365 (2023), https://doi.org/10.48550/arXiv.2305.04365.

https://doi.org/10.1017/S0009840X25000356 Published online by Cambridge University Press



Strongest OBJ of *do* by individual author

Latin	English	$-2\log\lambda$
Caesar		
obses	hostage	18.4
opera	service	11.9
suspicio	suspicion	2.2
facultas	faculty	1.8
signum	sign	1.5
Ovid		
osculum	kiss	8.5
velum	sail	5.9
munus	gift	3.5
signum	sign	2.6
Jerome		
potus	drink	16.6
esca	food	3.3
requies	rest	3.0
gloria	glory	2.4
terra	earth	1.6

FIGURE 7. Left: common objects for *faciō* by author; right, mock-up of a dynamic lexicon entry for the Latin verb *liberō*, 'to free' (figures from Bamman and Crane 2008)[32].

sixteen years after D. Bamman's pioneering work,[33] to implement such a dynamic lexicon for ancient Greek and Latin.

Bamman's work on the *Dynamic Lexicon* used treebank and alignment data to complement each other, and the goal from the start was always to provide both treebanks and alignments in a single reading environment. Integration of treebanks and alignments in a single reading environment has two main purposes. First and foremost, the goal of an integrated reading environment is to make primary sources intellectually more accessible. With an integrated environment, readers do not have to switch between reading environments for treebanks and alignments that are not aligned with each other (e.g. readers who start with treebanks cannot automatically move to translation alignments and vice versa).

Second, publications are only born-digital insofar as they can circulate across different projects and platforms. Reviews of born-digital publications necessarily not only include expository prose (such as that in this Profile of translation alignments and treebanks),

[32]These figures originally appeared as Table 9 and Figure 4 in D. Bamman and G. Crane, 'Building a dynamic lexicon from a digital library', *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (2008), pp. 11–20, https://doi.org/10.1145/1378889.1378892.

[33]The National Endowment for the Humanities funded the project 'The Dynamic Lexicon: Cyberinfrastructure and the automatic analysis of historical languages', PR-50013-08: 2007–2012.

https://doi.org/10.1017/S0009840X25000356 Published online by Cambridge University Press

but also demonstrate that third parties can in fact reuse those born-digital publications. In this, digital publication resembles not just copyright law (which requires original expression), but also patent law, which (at least in the United States) has the enablement requirement. A new invention must be described in terms that are sufficiently clear, concise and exact so that other people can use it. Likewise, a publication is not truly a digital publication unless other people can reuse it.

We look for several features when evaluating born-digital resources such as treebanks and translation alignments. (1) Data must be provided under a license that allows it to be uploaded and redistributed by a third party. In practice, those working on digital ancient Greek and Latin have converged around variations of the Creative Commons open licenses. (2) The data must be available in a sufficiently documented format so that other projects find it more effective to reuse existing data rather than to start over and build their own. Documentation includes transparency not only for the structures (e.g. Text Encoding Initiative TEI XML) but also for accuracy of the data (e.g. were the treebank annotations or alignments produced by automated systems or created/curated by human annotators?) (3) The data must have credits in a machine-actionable format that other projects can analyse and retain. This third element represents arguably the biggest current challenge to true digital publication. The Perseus Dependency Treebanks for Homer and Hesiod identify the annotators for each and every sentence. When F. Mambrini, an early contributor to the Perseus Treebanks and an expert in that form, created a version of the Perseus Homer Treebank in the Universal Dependency Framework, he did not include the sentence level credits because he had no standardised way of representing these credits in the UD format. Likewise, the latest edition of the GLAUx Treebanks includes both automatically and manually treebanked sentences. It identifies whether a sentence was manually or automatically annotated, but it does not (yet) have a mechanism to identify specific human annotators.

To establish these three features, we were able to use the Beyond Translation reading environment. We developed Beyond Translation with support from a variety of sources as a step towards a new version of the Perseus Digital Library. Beyond Translation did allow us to create the first (though surely not the last) reading environment that combined treebanks, translation alignments with other categories of annotation (metrical analysis, links to grammars and lexica, textual variants, named linking). The point is not that any one particular system has integrated this data, but rather that it was in fact practical to integrate alignments and treebanks.

CONCLUSION – TOWARDS THE OTHER AI, AUGMENTED INTELLIGENCE

An enormous amount of work remains to be done improving our ability to generate alignments and treebanks automatically, refining the results of automatic methods for manageable amounts, and on using new refined data to generate better models for automatic analysis. Nevertheless, we finally have in place the basic services that enable new forms of reading and make source texts in the original language intellectually accessible to new audiences. This addresses a critical challenge for linguistically constrained subjects such as ancient Greek and Roman culture where only a relative handful will develop mastery of ancient Greek or Latin, much less mastery of the research publications in languages such as French, German, Italian and Spanish (to mention only traditional European languages of publication). But if we have the basic tools, we are only now starting to understand how these tools are used and how users do and do not benefit. The most important research area for students of the past might not be how to

https://doi.org/10.1017/S0009840X25000356 Published online by Cambridge University Press

produce new articles and monographs but to learn how we can exploit digital methods to make the results of our work advance the intellectual life of society as a whole.

Tufts University GREGORY CRANE
gregory.crane@tufts.edu

Tufts University ALISON BABEU
alison.babeu@tufts.edu

University of Leipzig FARNOOSH SHAMSIAN
farnoosh.shamsian@uni-leipzig.de

https://doi.org/10.1017/S0009840X25000356 Published online by Cambridge University Press