BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **24**, No 4
Sofia • 2024 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2024-0041

The Browser-Based GLAUx Treebank Infrastructure: Framework, Functionality, and Future

Alek Keersmaekers, Frédéric Pietowski, Toon Van Hal, Mark Depauw

University of Leuven, Blijde-Inkomststraat 21, Leuven, Belgium

E-mails: alek.keersmaekers@kuleuven.be frederic.pietowski@kuleuven.be toon.vanhal@kuleuven.be mark.depauw@kuleuven.be

Abstract: *This paper presents the browser-based treebank infrastructure of GLAUx* (*the Greek Language AUtomated*). *This linguistic annotation project now has its integrated and user-friendly platform for exploring this data. After discussing the size and types of texts included in the GLAUx corpus, the contribution succinctly surveys the types of linguistic annotation covered by the project* (*morphology, lemmatization, and syntax*). *The emphasis of the contribution is on a description of the underlying SQL database structure and the search architecture. Infrastructure-related challenges faced by the GLAUx project are also discussed. Finally, the paper concludes with a discussion of future steps for the project, including additional functionality and expansion of the corpus.*

Keywords: *Ancient greek, Manual annotation, Automatic annotation, Treebank querying, Corpus linguistics, Search infrastructure.*

1. Introduction

Ancient Greek has long been the subject of study. Today, corpus linguistics has gained prominence in the study of modern languages, resulting in the availability of both manually and automatically annotated corpora. However, accessing annotated corpora for classical languages such as Ancient Greek is not as straightforward, despite the existence of numerous treebanks. To address this gap, the GLAUx project (Greek Language AUtomated) aims to consolidate existing manually annotated corpora and supplement them with a comprehensive automatically annotated corpus, as detailed by Keersmaekers [1]. The primary objective of this project is to offer researchers, educators, and students a user-friendly, web-based platform that allows them to effortlessly search and explore the entire corpus. It is essential to differentiate between the GLAUx annotated data and the GLAUx search infrastructure. A preliminary version of the search infrastructure was introduced on 9 March 2023, and can be accessed at <**www.glaux.be**>.

Keersmaekers [1] provides a description of the GLAUx data, which is via GitHub available online. While this paper also briefly outlines the size and types of text included in the searchable GLAUx corpus (note that not all available data are currently implemented in the search infrastructure) and surveys the types of linguistic annotation employed, the primary focus of this contribution is the architecture of the underlying database and search infrastructure. After giving an overview of the challenges faced by the project, the paper concludes with a discussion of our future plans, which include expanding the corpus and implementing additional functionality.

2. Texts in the GLAUx corpus

This subsection presents the types of texts available in the corpus, the sources, and the chronological range.

2.1. Which types of text?

In the realm of Ancient Greek texts, three main categories exist: literary, papyrological, and epigraphical texts. The literary texts, containing a wide range of (mostly) high-register texts including scientific, philosophical, and religious writings as well as literary genres in a more strict sense, like poems and narrative prose, have come down to us through the manuscript tradition. On the other hand, papyrological sources mostly scribbled on papyrus (even though graffiti is most often also regarded as papyrological sources), comprise everyday writing like letters and petitions. The more durable epigraphical texts, carved on stone, include decrees, epitaphs, and honorary inscriptions. The GLAUx search infrastructure is currently focused on analyzing literary texts, but the team plans to include papyrological texts soon. However, there is still a significant amount of work to be done in analyzing epigraphical texts.

2.2. Where do the GLAUx texts come from?

As the texts of the most authoritative source (the Thesaurus Linguae Graecae, henceforth: (TLG)) are not publicly available, the project is unable to make use of the most recent and optimal text editions. We made a selection of texts that are online available. Three sources offer more than 1M words (the Perseus Project; First Thousand Years of Greek and Wikisource).

2.3. What is the chronological range of GLAUx?

The ancient Greek literary corpus, which commences with the Homeric poems in the eighth century BC, is a complex and challenging entity to define due to the continued use of Greek well after the Byzantine Empire’s decline. The GLAUx corpus aims to include texts from the eighth century BC to the eighth century AD (the range of the Trismegistos project, see [2] and below) to incorporate both literary and non-literary works, spanning sixteen centuries of Greek. The current iteration of GLAUx focuses on literary texts up to the fourth century AD and uses a simplified genre classification system similar to that of the TLG. However, this classification will be refined in collaboration with Trismegistos to optimize automation and integration with other resources in the future.

3. Linguistic annotation

After briefly reviewing existing linguistic annotation projects related to Ancient Greek, we will survey the levels of annotation (see [1] for a more detailed description), starting with the relevant accomplishments to date and then discussing our plans for future work.

3.1. Annotation projects in the past

Several projects have manually annotated Greek texts for morphology, lemmas, (dependency) syntax, and occasionally semantics. We can single out the PROIEL project (more than 250K tokens [3]), the Ancient Greek Dependency Treebanks (AGDT; more than 550K tokens [4]), the Gorman trees (more than 300K tokens [5]), and the Pedalion project (more than 300K tokens [6]). Along with some more specific projects (see, e.g., [7] for a papyrological collection as well as [8] for earlier references), these annotators have collectively annotated about 1.5M tokens.

The GLAUx project heavily relies on these projects for two reasons. Firstly, we have integrated these annotations into our data, clearly marking the annotated sentences in our result list and giving credit to the original annotator on a sentence level. Secondly, we have made use of these annotations as training data to predict the annotation of raw texts.

3.2. Morphology

The part-of-speech classes are divided into nouns, adjectives, verbs, adverbs, pronouns, conjunctions, prepositions, numerals, articles, and interjections. The morphological annotation is consistent with the tag set of the Ancient Greek Dependency Treebank and includes person, number, tense/aspect, mood, voice, gender, case, and degree as attributes. For morphological and part-of-speech tagging, RFTagger was employed. This tagger uses decision trees and contextual probabilities, along with a morphological lexicon generated by the Morpheus morphological analysis tool to constrain the output.

Together with Wouter Mercelis, an ELECTRA-based model was recently designed (a BERT-like transformer-based language model), which will be applied in order to improve tagging results (a prototype has already been used in the current version). In addition, Alek Keersmaekers is currently expanding the morphological annotation with a derivational annotation layer, which will link complex morphological derivations to a stem (or root) on the one hand and a morphological pattern on the other hand, thus expanding linguistic research possibilities for end users.

3.3. Lemmatization

For the lemmatization process, Lemming was used, which applies formal, lemma, part-of-speech, morphology, and dictionary features. The accuracy of the lemmatization was initially 0.969 and was increased to 0.980 by relying on a Morpheus lexicon as a constraint. Lemmatization accuracy is generally high for Greek words due to the morphological complexity of the language. The lemmatization accuracy for poetic data is slightly lower than for prose data, with the accuracy ranging from 0.965 up to 0.975 for different poetic genres. The lemmas are generally consistent with the LSJ lexicon and the Morpheus codebase.

We are currently improving the quality of the lemmatization by removing inconsistencies and by applying a transformer-based pipeline for ambiguous cases (e.g., πείσομαι, which can be a future form of either πάσχω or πείθω).

3.4. Syntax

The GLAUx corpus was annotated with dependency information using AGDT (2.0) guidelines, based on the Prague Dependency Treebanks' format [9]. We made use of the Stanford Graph-Based Dependency Parser, which relies on character, token, and part-of-speech embeddings, with good results. The LAS score was 0.845 for papyri and ranged from 0.751 up to 0.881 for literary texts depending on the genre.

In the future, we aim to address some inconsistencies in our training data by improving our homogenization efforts. Next, we plan to transition to the Universal Dependencies annotation standard, which has wider support. Additionally, we intend to test new transformer-based parsers to improve the syntactic annotation.

3.5. Animacy

An annotation layer that was recently added is the animacy parameter. This annotation is restricted to nouns and allows users to identify the following categories: animal, body part, concrete, ethnonym, group, human, natural object, natural phenomenon, non-concrete, place, plant, time, and vehicle. The automatic prediction was based on machine learning (using gradient-boosted trees) on a dataset that was mainly annotated by students – we will report on the manual annotation and these experiments in a future publication. The overall accuracy is estimated to be 0.938, with F1 scores ranging from 0.75 (vehicle) up to 0.97 (human).

4. The database structure of GLAUx

The GLAUx infrastructure operates on an SQL-based architecture, motivated by the concerns raised by Onambélé et al. [10] regarding the computational cost of using XML files. To simplify, the MySQL architecture is visualized in Fig. 1.
The GLAUx SQL database consists of various tables with unique persistent identifiers. At the core is the word level with the GLAUx ID for each token, which also contains annotation information for part-of-speech, morphology, lemma, etc.

The unique feature of this architecture is its design which does not require sentences to be stored in a separate table. Rather, sentences are assembled dynamically through a special ID in the central table, which also encodes the order of words. As a result, the sentence can be easily reconstructed by querying all words associated with a given sentence ID in the appropriate sequence. This approach effectively eliminates redundancy as words are identified by unique numerical identifiers. Moreover, in the event of changes to the sentence, such as the addition or deletion of a word, the architecture ensures minimal complexity and easy modification.

This central table is linked with tables storing metadata information. The connection with Trismegistos Authors assures the availability of metadata for works and Greek authors.

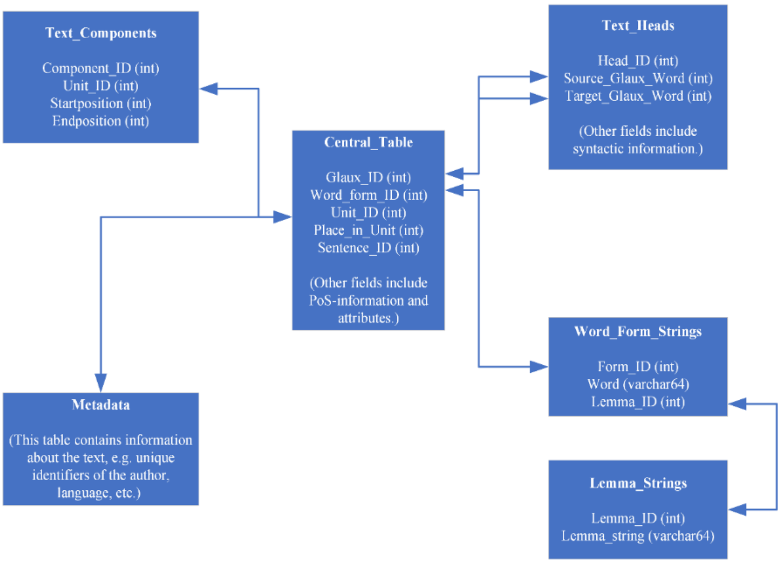
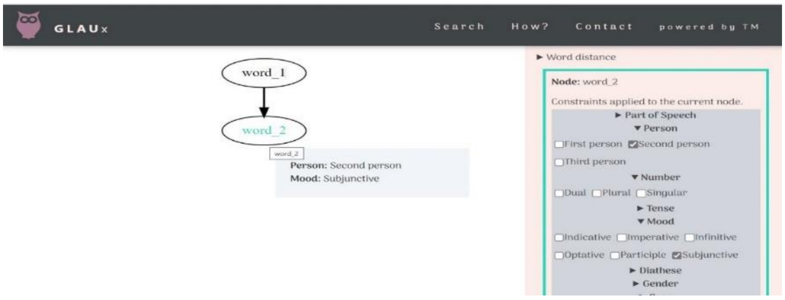


Fig. 1. A simplified visualization of the GLAUx search infrastructure

4.1. The search infrastructure architecture

The GLAUx database, as we have seen above, is a rich repository of information that is stored in a MySQL database. To make this information more accessible to users, Frédéric Pietowski has developed the GLAUx query infrastructure, which enables users to search for and retrieve data using a combination of PHP and JavaScript technologies. This system, which replaces the former offline DendroSearch tool by Keersmaekers et al. [6], consists of two main pages: The search page and the results page.



Person: Second person
Mood: Subjunctive

► Word distance

Node: word_2

Constraints applied to the current node.

► Part of Speech

▼ Person

□First person ☑Second person

□Third person

▼ Number

□Dual □Plural □Singular

► Tense

▼ Mood

□Indicative □Imperative □Infinitive

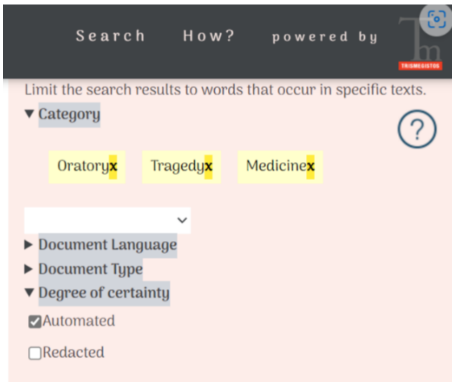
□Optative □Participle ☑Subjunctive

► Diathese

► Gender

Fig. 2. A screenshot of the GLAUx search page

The search page (Fig. 2) is designed with ease of use in mind, with two main sections that allow users to build hierarchical search queries with words (~nodes) and syntactic relations (~edges). The left-hand section presents a graphical environment where users can add or remove nodes from the graph. This is done by selecting the "Add descendant" or "Delete word & descendants" buttons that appear in the tooltip panel when the user clicks on a node. The resulting set of interconnected words and relations form a search instruction, which is visualized in the web environment. Users can interact with it by clicking on a node, which makes available a wide range of parameters to select from in the right-hand section. This allows users to specify both morphological features (e.g., noun, singular, accusative, and so on) and the syntactic relation (such as "object"). Finally, there is an input field that allows users to enter a linguistic "lemma" as a search term. First, this field is insensitive to both case and accentuation, with each unique word display linked to an ID. Only in the second stage, does the query become accentuation and case-sensitive, to optimize ease of use. Users can set a "distance constraint" to determine the maximum distance between elements in a set of nodes. Obviously, the distance must always be greater than or equal to the number of provided nodes. The right-hand section of the search page also enables users to add metadata constraints to the individual node specifications. Users can set a "distance constraint" to determine the maximum distance between elements in a set of nodes. Obviously, the distance must always be greater than or equal to the number of provided nodes.



Limit the search results to words that occur in specific texts.

▼ Category

Oratory x Tragedy x Medicine x

▶ Document Language

▶ Document Type

▼ Degree of certainty

☑Automated

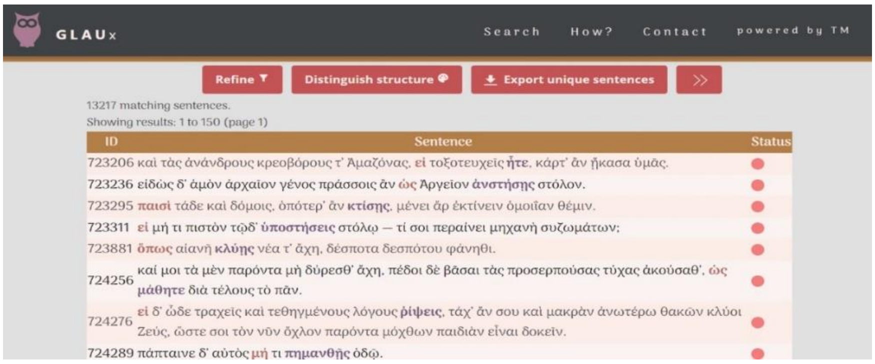
□Redacted

Fig. 3. A screenshot visualizing the possibilities for querying the metadata

As shown in Fig. 3, the set of available texts can be restricted based on genre and degree of certainty, which allows users to exclude automatically annotated texts. Future releases in collaboration with Trismegistos will add additional metadata constraints such as author, dialect, and century. In sum, this search page provides a visually intuitive way for users to build complex queries with syntactically related data. In the current version, it is not possible to make queries on word combinations that have no syntactic connection.

The results page (Fig. 4) displays the search results based on the user's query, with buttons for navigating, visualizing, and exporting search results. A PHP script processes the search request and returns a JSON response. After fetching the total number of matching sentences, the script displays the results in a table format.

Users may notice that certain sentences are duplicated in the search results. This occurrence arises from the fact that a particular construction may appear multiple times within a sentence. To provide clarity, the relevant terms are highlighted in bold within the search results. Moreover, users have the option to activate a special button that differentiates which words correspond to each query through the use of colors. Furthermore, the color of a bullet in the last column denotes whether the phrase was automatically or manually annotated. Finally, users can export the data in a CSV format.



13217 matching sentences.
Showing results: 1 to 150 (page 1)

ID	Sentence	Status
723206	καὶ τὰς ἀνάνδρους κρεοβόρους τ' Ἀμαζόνας, εἰ τοξοτευχεῖς ἦτε, κάρτ' ἂν ᾔκασα ὑμᾶς.	●
723236	εἰδὼς δ' ἀμὸν ἀρχαῖον γένος πράσσοις ἂν ὡς Ἀργεῖον ἀνστήσῃς στόλον.	●
723295	παισὶ τάδε καὶ δόμοις, ὁπότερ' ἂν κτίσῃς, μένει ἄρ ἐκτίνειν ὁμοιΐαν θέμιν.	●
723311	εἰ μή τι πιστὸν τῷδ' ὑποστήσεις στόλῳ — τί σοι περαίνει μηχανὴ συζωμάτων;	●
723881	ὅπως αἰανῆ κλύῃς νέα τ' ἄχη, δέσποτα δεσπότου φάνηθι.	●
724256	καί μοι τὰ μὲν παρόντα μὴ δύρεσθ' ἄχη, πέδοι δὲ βᾶσαι τὰς προσερπούσας τύχας ἀκούσαθ', ὡς μάθητε διὰ τέλους τὸ πᾶν.	●
724276	εἰ δ' ὧδε τραχεῖς καὶ τεθηγμένους λόγους ῥίψεις, τάχ' ἄν σου καὶ μακρὰν ἀνωτέρω θακῶν κλύοι Ζεύς, ὥστε σοι τὸν νῦν ὄχλον παρόντα μόχθων παιδιὰν εἶναι δοκεῖν.	●
724289	πάπταινε δ' αὐτὸς μή τι πημανθῇς ὁδῷ.	●

Fig. 4. A screenshot of the GLAUx results page

4.2. Challenges

In embarking on this project, we faced a multitude of challenges, which will be discussed in this section. Wherever possible, we will also suggest possible solutions, with a focus on infrastructure-related issues.

Keersmaekers [1] has already provided a comprehensive analysis of three main issues related to annotation, which we will only briefly summarize here. One of the challenges involves encoding multiple versions of ancient Greek texts. Ideally, different interpretations of a specific word or passage could be digitally represented by having several aligned versions of a given word form, allowing for automatic analysis of all these text variations. This would enable, for instance, the ability to control search results for the frequency of a particular construction by eliminating all disputed readings. However, due to the scarcity of openly accessible digital text editions, this remains an unattainable goal at the moment. Another challenge is the use of Natural Language Processing (NLP) techniques on a diverse corpus of Greek texts that spans multiple genres, periods, and dialects. Using NLP on out-of-domain data can lead to lower accuracy, while the impact of different factors such as genre, time period, and dialect on computational modeling accuracy may vary. To address this challenge, our team at KU Leuven has focused on diversifying the training data. Finally, Keersmaekers [1] emphasizes the influence of linguistic ambiguity stemming from historical language changes on data annotation. Given the extensive time span covered by the GLAUx corpus, it becomes challenging to develop a consistent annotation format for Greek linguistic constructions, as many of them undergo reanalysis over time.

Several challenges exist at the intersection of annotation and infrastructure. Among the existing manually annotated corpora, the PROIEL corpus stands out for its meticulous annotation, though it follows a different scheme than the Perseus format. Despite our conversion exercise, there may be some inconsistencies in the annotations caused by technical limitations. Additionally, an ongoing concern is the potential for human inconsistency between several annotators, particularly at the syntactic level. Such inconsistencies can have a significant impact on the annotation quality, making it a matter of importance to minimize them. In the future, AI approaches might assist us in detecting and correcting anomalies.

When it comes to the infrastructural challenges posed by the GLAUx project, we first need to ask ourselves a self-critical question – do we really need an additional tool to query annotated data? Vandeghinste and Augustinus [11], the creators of the GReTel treebank query infrastructure, rightly point out that there are already numerous linguistic treebanks available, each with their own query languages and exploration tools. While we appreciate the importance of projects that aim to provide a comprehensive structure to address the persistent problems of fragmentation (cf. initiatives like [12]), we are nevertheless currently compelled to use our own infrastructure. Our primary reasons for doing so are the need to optimize data regularly and allow users to contribute to annotation. Nonetheless, we are willing to offer our data to other initiatives, given that our underlying data system (SQL data) is highly adaptable. Data will also be provided in other formats (such as XML and CONLL).

While the use of unique identifiers such as the TLG reference system or the Trismegistos IDs enables interaction with other projects at the text level, the absence of unique identifiers at the sentence or word level poses a significant challenge. This means that it is difficult to exchange data and annotations at the sentence level, and even at the paragraph or chapter level, classifications remain a source of confusion for classical texts. GLAUx has taken the initiative to assign unique identifiers to each word, but a more sustainable solution is required to facilitate the organic growth of annotations and data exchange. We are currently exploring with Trismegistos and the new NIKAW project at KU Leuven on how best to model this.

Furthermore, the issue of data replication arises when updating the data regularly. How will users be able to replicate their research conducted based on an earlier version of the data? This is a significant concern that requires attention to ensure continuity and accessibility of research findings. Due to the high resource costs associated with providing the full previous version with each new data launch, we have opted to incorporate an inventory of changes into the existing infrastructure. Nevertheless, we still need to establish a system that would allow users to obtain replicable results.

The inclusion of the new animacy annotation layer (Section 3.5) introduces a new complication. Until recently, we annotated complete sentences at the levels of morphology, syntax, and lemma. Fully checked sentences show a green circle, while automatically analyzed sentences show a red one, as shown in Fig. 4. Animacy, however, is independent of this approach. It applies solely to nouns and does not correlate with sentence-level annotations. Therefore, sentences manually annotated for animacy are not automatically accompanied by other annotation layers, and vice-versa. We need to find a way to signal this divergence in the tool.

Lastly, sustainability poses a significant challenge. The GLAUx search service is resource-intensive, with millions of records, and possibly heavy queries from users.

Maintaining all the Javascript and PHP applications demands special expertise. While the GLAUx data is available for free on GitHub, it may be necessary to reserve some of the queries for paying Trismegistos users to ensure the long-term viability of the project.

5. Future steps

5.1. Annotation

As we continue to refine our natural language processing techniques and collect more training data, we aim to periodically update the GLAUx data.

In addition to the existing annotation layers (syntax, morphology, lemma), GLAUx will also offer semantic role annotation. To this aim, Keersmaekers [13] expanded and revised the semantic role set used in the Pedalion project [14] and made it compatible with frameworks used for other languages, such as FrameNet. Multiple roles are distinguished, including agent, beneficiary, and recipient. First tests with a Random Forest classifier developed by Keersmaekers achieved an accuracy ranging from 0.687 up to 0.838 across different text types, despite having a relatively small number of training examples. In addition, we are also experimenting with automated Word Sense Disambiguation [15].

5.2. Additional texts

In the next release, we anticipate integrating the papyri into this infrastructure, the annotated data that Alek Keersmaekers has shared on the GitHub platform (**https://github.com/alekkeersmaekers/dukenlp**). As it stands, GLAUx's current launch lacks literary texts from the fourth century AD onwards. In the next iteration, we hope to expand the breadth of our data to include more recent sources. The inclusion of epigraphic texts, however, presents its own unique set of challenges, which we will address in due course. Moreover, we remain open to the possibility of including smaller corpora of alternative languages in future releases. For instance, we are considering the CEIPoM corpus, curated by Reuben Pitts, which includes meticulously manually crafted treebanks of Oscan, Umbrian, Messapian, Venetic, and Old Latin [16].

5.3. Functionality

While it may seem obvious to include a text browsing function in GLAUx, we have chosen to forgo this option in favor of using the Scaife viewer (for a succinct discussion, see [17]), which offers the ability to integrate GLAUx data and allows readers to easily obtain additional information by hovering over words and viewing the entire syntactic tree. If integration within the Scaife viewer would take a long time, a limited browsing system can be implemented within GLAUx (the infrastructure is in place), but we would prefer to limit expansion rather than increase it.

In a future version of GLAUx (in collaboration with Trismegistos Authors), users will be able to search for texts based on century, author, and genre, as well as on isolated word sequences.

We also aim to involve users in the annotation process by allowing them to correct and validate automatically analyzed sentences, and they will be duly credited for their contributions. The infrastructure for this is already largely in place.

Connecting the annotation data to the modular Pedalion grammar [14] is a significant challenge that we are currently tackling. Our aim is to link nodes in the text to corresponding nodes in the grammar, which we believe will be made possible after the integration of semantic roles. As part of this effort, we are also revising the modular syntax to ensure everything is properly aligned.

5.4. Linking with Trismegistos and other projects

GLAUx was built in close cooperation with Trismegistos, the platform for the study of the ancient Western world (800 BC – AD 800). Trismegistos Words, an online search engine for the lexicon and morphology of Greek papyrological texts, can be seen as a pilot project for GLAUx [18]. We are currently also exploring further integrating GLAUx with Trismegistos Authors so that search functionality can be enhanced with more metadata. Finally, Trismegistos is developing an engine to link references in secondary literature to metadata and the actual text of the sources themselves. GLAUx is also an important partner for this project.

In conclusion, we must consider how this project relates to other annotation efforts. This includes not only Ancient Greek projects but also other historical initiatives, which are increasingly numerous. Different goals and resources often put these projects on varied technical platforms, complicating intercommunication. For an early overview, see Chapter 8 of M. Piotrowski's *Natural Language Processing for Historical Texts* [19]. His book also emphasizes the need for more theory-building in the field bridging digital humanities and NLP approaches.

Acknowledgments**:** The GLAUx project had its beginnings in a spontaneous and organic manner, emerging against the backdrop of Alek Keersmaekers's doctoral project (2016-2020), the Pedalion educational project, and the Trismegistos project. Since 2019, the Research Foundation Flanders (FWO) has provided funding for the project, specifically through "Towards a breakthrough in the automated parsing of Ancient Greek papyrological and literary texts? Extending and refining the available training data (1523619N)" and "Language and Ideas: Towards a New Computational and Corpus-based Approach to Ancient Greek Semantics and the History of Ideas (G052021N)". We extend our heartfelt gratitude to the many manual annotators of ancient Greek texts, whose dedicated efforts have made this project possible.

References

1. Keersmaekers, A. The GLAUx Corpus: Methodological Issues in Designing a Long-Term, Diverse, Multi-Layered Corpus of Ancient Greek. – In: Proc. of 2nd International Workshop on Computational Approaches to Historical Language Change 2021. Association for Computational Linguistics, 2021, pp. 39-50.
2. Depauw, M., T. Gheldof. Trismegistos: An Interdisciplinary Platform for Ancient World Texts and Related Information. – In: Proc. of International Conference on Theory and Practice of Digital Libraries, Berlin, Heidelberg, Springer, 2013, pp. 40-52.
3. Haug, D., M. Jøhndal. Creating a Parallel Treebank of the Old Indo-European Bible Translations. – In: Proc. of 2nd Workshop on Language Technology for Cultural Heritage Data, 2008, pp. 27-34.

4. Bamman, D., F. Mambrini, G. Crane. An Ownership Model of Annotation: The Ancient Greek Dependency Treebank. – In: Proc. of Eighth International Workshop on Treebanks and Linguistic Theories (TLT 8), 2009, pp. 5-15.
5. Gorman, V. Dependency Treebanks of Ancient Greek Prose. – Journal of Open Humanities Data, Vol. **6**, 2020, No 1.
6. Keersmaekers, A., W. Mercelis, C. Swaelens, T. Van Hal. Creating, Enriching and Valorising Treebanks of Ancient Greek: The Ongoing Pedalion-Project. – In: Proc. of 18th International Workshop on Treebanks Association for Computational Linguistics (ACL) and Linguistic Theories (TLT, SyntaxFest), Paris, 2019.
7. Vierros, M., E. Henriksson. PapyGreek Treebanks: A Dataset of Linguistically Annotated Greek Documentary Papyri. – Journal of Open Humanities Data, Vol. **7**, 2016.
8. Celano, G. The Dependency Treebanks for Ancient Greek and Latin. – In: M. Berti, Ed. Digital Classical Philology. Berlin, Boston, De Gruyter, 2019, pp. 279-298.
9. Bamman, D., G. Crane. The Ancient Greek and Latin Dependency Treebanks. – In: Language Technology for Cultural Heritage. Berlin, Heidelberg, Springer, 2011, pp. 79-98.
10. Onambélé, Chr., M. Kopp, M. Passarotti, J. Mírovskỳ. Converting Latin Treebank Data into an SQL Database for Query Purposes. – In: Proc. of 2nd International Conference on Digital Access to Textual Cultural Heritage, 2017, pp. 117-122.
11. Vandeghinste, V., L. Augustinus. Making a Large Treebank Searchable Online: The SoNaR Case. – In: Proc. of LREC2014 2nd Workshop on Challenges in the Management of Large Corpora (CMLC-2), ELRA, Paris, 2014, pp. 15-20.
12. Krause, Th., A. Zeldes. ANNIS3: A New Architecture for Generic Corpus Query and Visualization. – Digital Scholarship in the Humanities, Vol. **31**, 2016, pp. 118-139.
13. Keersmaekers, A. Automatic Semantic Role Labeling in Ancient Greek Using Distributional Semantic Modeling. – In: Proc. of the 1st Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA'20). European Language Resources Association (ELRA), Marseille, 2020, pp. 59-67.
14. Van Hal, T., Y. Anné. Reconciling the Dynamics of Language with a Grammar Handbook: The Ongoing Pedalion Grammar Project. – Digital Scholarship in the Humanities, Vol. **32**, 2016, No2, pp. 448-454.
15. Popov, A. Neural Network Models for Word Sense Disambiguation: An Overview. – Cybernetics and Information Technologies, Vol. **18**, 2018, No 1, pp. 139-151.
16. Pitts, R. J. Corpus of the Epigraphy of the Italian Peninsula in the 1st Millennium BCE (CEIPoM). – Journal of Open Humanities Data, Vol. **8**, 2022.
17. Muellner, L. The Free First Thousand Years of Greece. – In: M. Berti, Ed. Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution. Berlin, De Gruyter, 2019, pp. 55-89.
18. Keersmaekers, A., M. Depauw. Bringing Together Linguistics and Social History in Automated Text Analysis of Greek Papyri. – A. Novokhatko, Ed. Digital Classics III: Re-Thinking Text Analysis (Classics@). Washington DC, Center for Hellenic Studies, 2021.
19. Piotrowski, M. Natural Language Processing for Historical Texts. Synthesis Lectures on Human Language Technologies. Cham, Springer, 2012.

Received: 11.11.2024; Accepted: 21.11.2024. Fast-track.