From Lemmas to Links: A Lemma Bank for Ancient Greek

Colin Swaelens†, Francesco Mambrini*, Marco Passarotti*
†Language & Translation Technology Team & Dpt. of Linguisitcs; Ghent University
*CIRCSE, Università Cattolica del Sacro Cuore (Milano)
colin.swaelens@ugent.be, francesco.mambrini@unicatt.it, marco.passarotti@unicatt.it

Abstract

This paper presents the initial construction of a lemma bank for Ancient Greek, developed according to the Linked Data principles. The need for interoperable linguistic infrastructures capable of supporting interoperability among historical variation, divergent annotation practices, and resource-specific lemmatisation conventions was highlighted by the increasing availability of digital linguistic resources. This lemma bank is developed as the core component of the Linking Greek knowledge base and inspired by the architecture of the LiLa project for Latin. The proposed lemma bank adopts a descriptive, lemma-centric approach that preserves alternative canonical forms and dialectal variation while enabling consistent linking across lexical and semantic resources. Its population combines data extracted from the Ancient Greek WordNet and the Liddell–Scott–Jones lexicon, integrating semantic structure, part-of-speech information, and lexicographically encoded gender assignment. Additional normalisation steps, including the rule-based correction of closed-class part-of-speech categories and harmonisation to the Universal Dependencies tagset, were applied to improve consistency and computational usability. The resulting dataset provides a foundation for interlinking corpora, lexica, and NLP tools for Ancient Greek within a Linked Open Data framework.

Keywords: Linked Open Data, Ancient Greek, Linguistic Resources

1. Introduction

The rapid growth of the field of ancient language processing (ALP) can be attributed, in part, to the increasing availability of both lexical and textual resources (Sommerschield et al., 2023). For Ancient Greek, this development has resulted in a rich landscape of digital resources. These include textual resources like the Perseus Digital Library[1] and the First1KGreek Project[2], syntactically annotated corpora such as the Ancient Greek Dependency Treebanks (Bamman and Crane, 2011) and the Pedalion Trees (Keersmaekers et al., 2019), semantic resources such as the Ancient Greek WordNet (Bizzoni et al., 2014), as well as lexical reference works, among which Liddell-Scott Jones (LSJ)[3].

These resources, however, adopt different annotation strategies, display a variety of data formats, and employ incompatible label sets: an issue well attested for many other languages as well. For Ancient Greek, this fragmentation is further shaped by historical and linguistic factors: prior to the emergence and subsequent diffusion of the Koine (4th c. BC), Greek did not constitute a single, standardised language, but rather a collection of regional dialects, some of which differ substantially in their phonology (Example 1a), morphology (Example 1b), and lexicon (Example 1c).

(1) a. Attic θάλαττα *thalatta* vs.
Ionic θάλασσα *thalassa* (sea)

b. Attic εἶναι *einai* vs.
Doric ἔμμεν *emmen* (to be)

c. Attic αὐτόν *auton* vs.
Ionic μιν *min* (acc. sg. 3rd pers. pron.)

This linguistic variation amplifies the heterogeneity already inherent in the existing resources, which negatively affects their use and hinders interoperability. The LSJ, for example, records both variants of Example 1a as entries, while the Greek–Dutch reference dictionary (Sluiter et al., 2024) consistently selects the Attic variant as headword. This may lead to artificially low scores in evaluation settings—for instance, when lemmatisation systems are penalised for producing a dialectal variant that is linguistically valid, yet absent from the chosen gold standard, as described in Swaelens et al. (2024).

To address issues of interoperability across resources, the principles of Linked Open Data (LOD) have been increasingly applied to linguistic data. Within this context, a lively research community has emerged around Linguistic Linked Open Data (LLOD), which promotes the publication of linguistic resources in accordance with Linked Data standards and Semantic Web technologies. See, for instance, the NexusLinguarum COST Action (CA18209), which fostered collaboration in the field

[1] https://www.perseus.tufts.edu/
[2] https://opengreekandlatin.github.io/First1KGreek/
[3] https://github.com/helmadik/LSJLogeion

Proceedings of LT4HALA 2026 @ LREC 2026, pages 106–111.

of LLOD. By representing linguistic information as interlinked entities identified by persistent URIs, LOD enables explicit connections between otherwise distinct datasets while preserving their internal annotation choices.

One of the most fully developed instantiations of the LOD principles is the LiLa (Linking Latin) project (Passarotti et al., 2020), which builds a linked data-based knowledge base of linguistic resources and natural language processing tools for Latin. The LiLa knowledge base's principal component is the lemma bank. Lemmas provide a stable point of reference across resource types: they occur as entries in lexical resources, their inflected forms appear as tokens in textual resources, and they constitute the unit over which many NLP tools operate. This makes them a natural anchoring unit for interoperable linguistic modelling.

Building on the experience of LiLa and reusing its underlying architecture, the Linking Greek project has initiated the construction of a knowledge base of interoperable linguistic resources for Ancient Greek as linked data. This paper describes the development of its principal component: a collection of Greek lemmas, hereafter referred to as the *lemma bank*. The lemma bank is designed to serve as an anchoring layer between word occurrences in corpora and their corresponding entries in lexical resources, thereby enabling consistent linking across heterogeneous datasets within the knowledge base.

2. Background

Building on the foundations of the Semantic Web (Berners-Lee et al., 2001), the Linked Data paradigm is organised around a set of four core principles that enable the data to be identified, queried and semantically interpreted (Berners-Lee, 2006):

1. Use Uniform Resource Identifiers (URIs) to name "things"
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to URIs so that they can discover more things

These principles have been successfully applied in a number of linguistic knowledge bases, illustrating how Linked Data can be used to model linguistic resources in practice. The LiLa knowledge base for Latin is a prominent example, as it represents linguistic objects (occurrences of words in texts, lexical entries in dictionaries, parts-of-speech, etc.) as URIs that can be accessed via HTTP. By explicitly linking lemmas to textual occurrences and annotations, LiLa demonstrates how heterogeneous datasets can be integrated while preserving their internal structure and annotation schemes. In addition, LiLa interlinks several lexical resources, including an etymological dictionary (de Vaan, 2008), thereby enabling connections between Latin lemmas and their Proto-Indo-European roots[4]. This integration illustrates how Linked Data principles facilitate cross-linguistic linking.

Similar approaches have been adopted for other (historical) languages. For Old Irish, a Linked Data knowledge base has been developed that converts the GoldLex lexical dataset into a lemma bank (Fransen et al., 2024), focussing primarily on nominal morphology and demonstrating the applicability of LLOD principles to historically less-resourced languages. A comparable application of the Linked Data paradigm can be found in LiIta, a knowledge base for Italian that already initiated its core component, the lemma bank (Litta et al., 2024)[5]. Together, these resources illustrate how Linked Data principles can be applied to linguistic data at different levels of scope and maturity, with lemma-centric modelling emerging as a recurring design pattern across languages.

3. The Greek Lemma Bank

Initiating the Linking Greek knowledge base with a lemma bank is a natural design choice, already motivated by the architecture of its Latin counterpart (Passarotti et al., 2020), in which lemmas function as anchoring units across different types of lexical, linguistic and textual resources.

In this approach, lemmatisation is treated as a shared organisational layer across different kinds of resources. Lemmas serve as indexing units in dictionaries and thesauri, enable the aggregation of semantically related entries, and facilitate lexical access to corpora, an especially crucial function for morphologically rich languages. In Ancient Greek, for example, a single verbal lemma can represent up to 242 forms of a verb, the participle forms not included.

On this basis, the lemma emerges as the most productive interface between lexical resources, annotated corpora, and NLP tools.

3.1. Resources

To construct the initial version of the Ancient Greek lemma bank, three complementary resources were considered: the lemma list of the morphological parsing tool Morpheus (Crane, 1991), the LSJ,

[4] https://lila-erc.eu/data-page/.
[5] https://www.liita.it

Resource	Number	Primary focus
Morpheus	45,868	Morphology
LSJ	116,924	Lexicography
AGWN	112,513	Semantics

Table 1: Overview of the primary resources considered for the construction of the Greek lemma bank.

and the Ancient Greek Word Net (Bizzoni et al., 2014). These resources represent distinct perspectives on the Greek lexicon, reflecting respectively a parser-oriented, a lexicographic, and a semantic approach to lemma representation. Each resource contributes a different lemma inventory, yet none of them, in isolation, provides a sufficient basis for a lemma bank intended to support interoperability across heterogeneous linguistic resources. To construct the initial version of the Ancient Greek lemma bank, three complementary resources were considered: the lemma list of the morphological parsing tool Morpheus (Crane, 1991), the LSJ, and the Ancient Greek Word Net (Bizzoni et al., 2014). These resources represent distinct perspectives on the Greek lexicon, reflecting respectively a parser-oriented, a lexicographic, and a semantic approach to lemma representation. Each resource contributes a different lemma inventory, yet none of them, in isolation, provides a sufficient basis for a lemma bank intended to support interoperability across heterogeneous linguistic resources.

While alternative approaches based on lemmatised corpora or treebanks were considered, we opted to prioritise lexical resources. In annotated corpora like Diorisis (Vatri and McGillivray, 2018) and Opera Graeca Adnotata (Celano, 2024), lemma assignment is often context-dependent and may vary according to syntactic interpretation, which makes it less suitable as a stable, canonical reference layer. For the purposes of constructing a lemma bank as an interoperability layer, a consistent and context-independent representation of lemma identity is required. Lexical resources, by contrast, provide a more controlled and abstraction-oriented representation of lemmas, which better aligns with the objectives of a lemma-centric Linked Data infrastructure.

Morpheus' lemma list provided a useful starting point, since it records not only lemmas but also associated linguistic information like part-of-speech and, for nouns, gender, in addition to the dialectal labelling, as noted above. Nevertheless, its coverage is comparatively limited and therefore insufficient as the sole backbone of the knowledge base (Table 1).

Coverage was not a limitation for the LSJ, whose lexicographic content is both extensive and reliable. The dictionary also records a substantial amount of dialectal variation (cf. Example 1); however, extracting this information automatically proved challenging, mainly because the relevant relations are encoded in heterogeneous and, at times, inconsistent ways. For phonological variation (cf. Example 1a), the entry for θάλαττα *thalatta* ('sea') explicitly states that the headword and the related verbal and adjectival forms, constitutes the Attic counterpart of θάλασσα (Example 2a). Yet comparable alternations are elsewhere encoded via cross-references using *v.*, as in τάττω *tattō* 'to put' *v.* τάσσω (Example 2b). Importantly, the same *v.* relation is also used for derivational or morphosyntactic links that are not dialectal variants at all, such as χάριν (an adverbial use of the accusative of χάρις) *v.* χάρις (Example 2c). As a consequence, it was not possible to automatically extract and link these forms to one another.

(2) a. θάλαττα, θαλαττεύω, θαλάττιος, etc., Att. for θάλασσα, etc.

b. τάττω v. τάσσω

c. χάριν v. χάρις

The Ancient Greek Wordnet provides a large-scale, systematically structured lemma inventory that approaches the coverage of the *LSJ*. In line with the WordNet paradigm, AGWN is organised around synsets representing lexicalised concepts, with lemmas functioning as the linguistic realisations of semantic senses rather than as dialectally normalised headwords. As a result, dialectal and phonological variation is not represented consistently: while AGWN includes both θάλασσα *thalassa* and θάλαττα *thalatta*, it only lists τάσσω *tassō* and not its Attic counterpart τάττω *tattō*. AGWN further provides part-of-speech and morphological information to support computational processing, although its annotation scheme does not fully align with Universal Dependencies conventions. Despite these limitations, AGWN offers a conceptually clean and explicitly structured view on lemma identity, which complements the lexicographic orientation of *LSJ* and the parser driven design of Morpheus.

3.2. The ontology of Ancient Greek canonical forms

Alternative canonical solutions adopted by different resources are modelled using the property `lila:lemmaVariant`, a symmetric relation defined in the LiLa ontology that connects different inflected forms of the same word that can be chosen for lemmatisation. The ontology also retains the class `lila:Hypolemma`, introduced in LiLa to model forms that may function as lemmas and are members of the regular inflectional paradigm of another lemma, with a different part of speech (e.g. participles: ADJ, members of verbal paradigms).

While this modelling option is preserved for conceptual compatibility, current evidence from Greek corpora suggests that participles are not typically used as canonical lemmas; consequently, no instances of `lila:Hypolemma` are instantiated in the present version of the Greek lemma bank. Orthographic differences that do not affect inflectional behaviour are represented separately using the `ontolex:writtenRep` property provided by OntoLex-Lemon. The ontology distinguishes the following set of core entities that capture the abstraction levels required to model Ancient Greek canonical forms and their variation:

- **Lemma**: the index of all inflected forms that is conventionally identified as canonical. For example, the lemma γράφω *graphō* (to write) indexes inflected forms such as γράφει *graphei* (A. ind. pres. 3rd. sg.) and ἔγραψεν *egrapsen* (A. ind. aor. 3rd. sg.).
- **Lemma Variant**: the symmetric property to link multiple forms that could be used as lemma for a specific token. An example could be the co-existence of the active and mediopassive voice of verbs in a lexicon, such as γράφω *graphō* and γράφομαι *graphomai*.
- **Written Representation**: the different spellings or peculiar inflections canonical forms can have. An example we already discussed above is the alternation of θάλασσα *thalassa* and θάλαττα thalatta. This property is used only when the variation in the realisation of the lemma affects only the orthography of a form, provided that its morphology is not altered.

These entities are connected through a small set of explicit properties that formalise how canonical forms, their variants, and their representations interact within the lemma bank:

- **lila:hasLemma**: links a lemma to its (inflected) word forms occurring in textual resources.[6]
- **lila:lemmaVariant**: a symmetric property linking lemma variants that represent alternative canonical solutions for the same lexical item.[7]
- **ontolex:writtenRep**: associates a lemma with its orthographic realisations.[8]

The population of the Greek lemma bank was initiated from the Ancient Greek WordNet, which served as the primary source for lemma extraction. Using the AGWN API index call, we retrieved the full set of lemmas together with all associated part-of-speech analyses. This proved particularly valuable, as a single written form may receive multiple part-of-speech assignments, which could be directly modelled as lemma variants within the ontology. At the same time, inspection of the AGWN data revealed inconsistencies in part-of-speech annotation, plausibly related to its largely automatic annotation pipeline; for instance, a number of prepositions are tagged as pronouns.

To improve the reliability of the part-of-speech information, an additional normalisation step was introduced for closed-class words. Using the lemma list provided by Morpheus as a reference inventory, closed-class items such as prepositions, conjunctions, particles, and pronouns were identified and used to correct inconsistent part-of-speech assignments in the extracted data. This rule-based correction was applied prior to mapping the resulting PoS labels to the Universal Dependencies tag set.

A further limitation of AGWN is the absence of gender information for nominal lemmas. To address this gap, information from the *Liddell–Scott–Jones* lexicon was integrated. Despite its heterogeneous and lexicographically oriented structure, LSJ consistently characterises nominal entries by means of the definite article, which allows both the unambiguous identification of nouns and the reliable extraction of their grammatical gender. This information was therefore combined with the lemma and part-of-speech data derived from the AGWN.

3.3. Resulting Data Set

The resulting version of the Greek lemma bank comprises *100,280* lemmas. Of these lemmas, *41,173* are nominal, for which grammatical gender could be assigned on the basis of LSJ evidence. The distribution of part-of-speech categories reflects the predominance of open-class items, with verbs and nouns accounting for the majority of entries, while closed-class categories were normalised following the procedure described above. All part-of-speech labels were harmonised to the Universal Dependencies tag set. Table 2 summarises the distribution of lemmas across Universal Dependencies part-of-speech categories in the current version of the Greek lemma bank.

One remaining issue concerns the distinction between common nouns and proper nouns, in particular personal names, which is not systematically captured in the current resources. Addressing this distinction is left for future work and may benefit from the integration of gazetteers or named entity resources.

[6] http://lila-erc.eu/ontologies/lila/hasLemma

[7] http://lila-erc.eu/ontologies/lila/lemmaVariant

[8] http://www.w3.org/ns/lemon/ontolex#writtenRep

UD Part-of-Speech	Number of lemmas
NOUN	41,173
ADJ	28,827
VERB	27,573
ADV	2,509
ADP	73
SCONJ	49
PART	32
PRON	21
NUM	14
CCONJ	5
INTJ	4

Table 2: Distribution of lemmas in the Greek lemma bank across Universal Dependencies part-of-speech categories.

4. Conclusion

This paper presented the initial construction of a lemma bank for Ancient Greek, developed as the core component of the Linking Greek knowledge base and explicitly inspired by the architecture of the LiLa project for Latin. By integrating data from heterogeneous lexical and semantic resources within a shared ontology, the lemma bank provides a unified, lemma-centric layer that supports interoperability across corpora, lexica, and NLP tools, while preserving resource-specific lemmatisation practices. The current version of the lemma bank is populated primarily from the Ancient Greek WordNet and the *Liddell–Scott–Jones* lexicon, combining semantic structure, part-of-speech information, and nominal gender assignment. Additional normalisation steps, including the rule-based correction of closed-class part-of-speech categories and harmonisation to the Universal Dependencies tag set, were introduced to improve consistency and computational usability.

Several directions for future research naturally follow from this work. First, the distinction between common nouns and proper nouns, in particular personal names, remains to be addressed and could benefit from the integration of gazetteers or named entity resources. Second, extending the coverage of dialectal and phonological variation, as well as refining the modelling of inflectional and derivational relations, would further enhance the descriptive power of the lemma bank. Finally, the integration of additional textual resources and the publication of the lemma bank as Linked Open Data will enable broader reuse and facilitate cross-linguistic linking within the wider LOD ecosystem.

5. Acknowledgements

This publication is based upon work from COST Action CA23147 GOBLIN – Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology, https://www.cost.eu).

6. Bibliographical References

References

David Bamman and Gregory Crane. 2011. The ancient greek and latin dependency treebanks. In *Language Technology for Cultural Heritage*, pages 79–98, Berlin, Heidelberg. Springer Berlin Heidelberg.

Tim Berners-Lee. 2006. Linked data – design issues.

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. Web semantic. *Scientific American*, 284(5):34–43.

Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini, and Gregory Crane. 2014. The making of Ancient Greek WordNet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1140–1147, Reykjavik, Iceland. European Language Resources Association (ELRA).

Giuseppe A. Celano. 2024. Opera graeca adnotata: Building a 34m+ token multilayer corpus for ancient greek.

Gregory Crane. 1991. Generating and parsing classical greek. *Literary and Linguistic Computing*, 6(4):243–245.

Michiel de Vaan. 2008. *Etymological Dictionary of Latin: and the other Italic Languages*. Brill, Leiden, The Netherlands.

Theodorus Fransen, Cormac Anderson, Sacha Beniamine, and Marco Passarotti. 2024. The MOLOR lemma bank: a new LLOD resource for Old Irish. In *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*, pages 37–43, Torino, Italia. ELRA and ICCL.

Alek Keersmaekers, Wouter Mercelis, Colin Swaelens, and Toon Van Hal. 2019. Creating, enriching and valorizing treebanks of Ancient Greek. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 109–117, Paris, France. Association for Computational Linguistics.

Eleonora Litta, Marco Passarotti, Paolo Brasolin, Giovanni Moretti, Valerio Basile, Andrea Di Fabio,

and Cristina Bosco. 2024. The lemma bank of the LiITA knowledge base of interoperable resources for Italian. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 517–522, Pisa, Italy. CEUR Workshop Proceedings.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin. *Studi e Saggi Linguistici*, 58(1):177–212.

Ineke Sluiter, Lucien van Beek, Ton Kessels, and Albert Rijksbaron. 2024. *Woordenboek Grieks/Nederlands*. Amsterdam University Press.

Thea Sommerschield, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. Machine learning for ancient languages: A survey. *Computational Linguistics*, 49(3):703–747.

Colin Swaelens, Pranaydeep Singh, Ilse de Vos, and Els Lefever. 2024. Lemmatisation of medieval Greek: Against the limits of transformer's capabilities? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10293–10302, Torino, Italia. ELRA and ICCL.

A. Vatri and B. McGillivray. 2018. The diorisis ancient greek corpus: Linguistics and literature. *Research Data Journal for the Humanities and Social Sciences*, 3(1):55 – 65.