Analysis of Anachronistic Lemmas and Semantic Fields in Ancient Greek WordNet

Gianluca Scatigno
Alma Mater Studiorum Università di Bologna, Bologna - Italy
Piazza San Giovanni in Monte, 2
gianluca.scatigno@unibo.it

Abstract

The Ancient Greek WordNet is a valuable resource for classical studies, but its accuracy is compromised by the presence of anachronistic semantic fields and lexical entries. This study conducts a comprehensive analysis to identify and categorize these anachronisms within the WordNet framework. The study systematically reviews and critiques semantic and lexical elements that are misaligned with the linguistic, cultural, and historical context of the ancient Greek world.

1 Ancient Greek WordNet

Over the past decade, there has been a significant proliferation of semantic-lexical resources for various languages, both modern and ancient, based on the WordNet relational model. This growth can be attributed not only to the crucial role of WordNets in linguistic studies but also to advances in artificial intelligence, particularly in Natural Language Processing (NLP), where many tasks take benefit of these databases[1].

The need to develop such resources has been especially critical for ancient languages like Latin and Ancient Greek[2]. For the latter, the Ancient Greek WordNet was created, and its structure will be briefly described here.

The Ancient Greek WordNet (AGWN) was modeled after the Princeton WordNet (PWN), following the same structure of synsets—groups of synonymous nouns, verbs, adjectives, and adverbs—organized in a hierarchical framework. This relational model captures semantic relations like hypernymy and hyponymy[3].

In AGWN, hyponymy and hypernymy are represented through semfields, which group synsets into broad macrocategories. These are defined using a numbering system based on the Dewey Decimal Classification (DDC), a general knowledge organization system commonly used in libraries and archives. The well-defined categories and extensive network of relationships within the DDC make it universally applicable across languages[4].

The backbone structure of AGWN consists of ten numbered classes, subdivided into ten divisions, which are further broken down into more specific sections. For example, class 100 (Philosophy & Psychology) includes divisions such as 110 (Metaphysics) and 120 (Epistemology), which can be further detailed into more specific categories. AGWN adopts this nested hierarchical structure to provide an adequate conceptual coverage, along with robust data organization[5].

Despite its strengths, AGWN has limitations in representing Ancient Greek, particularly due to anachronisms. Occasionally, lemmas may be associated with definitions that are completely out of context for classical antiquity.

These issues primarily stem from the data sources and the construction process of AGWN. Firstly, relying on Greek-English bilingual dictionaries may fail to capture the subtleties of ancient Greek semantics and polysemy, resulting in errors like the one previously mentioned. Additionally, the alignment with Princeton WordNet (PWN), which is based on contemporary English, may introduce anachronisms and errors that undermine semantic coherence[6].

Moreover, the classification system used in AGWN reflects modern knowledge organization schemes, which may not align with historical frameworks[7]. The Dewey system is not free from

[1] Bizzoni et al. 2014, 1140.
[2] Bizzoni et al. 2014, Minozzi 2009.
[3] Fellbaum 2005, 665-667.
[4] Bentivogli et al. 2014, 96.
[5] The Dewey Decimal Classification has been employed within the WordNet Domains Hierarchy project, which aimed to reorganize WordNet content across languages (vd. Bentivogli et al. 2014).
[6] Bizzoni et al. 2015, 47.
[7] Dewey published the classification in 1876 and it has

bias, rooted in the Eurocentric and colonial context in which it was developed. This, combined with Dewey's strong Christian orientation, has influenced the classification approach, leading to a predominant representation of Christianity and a neglect of other religions, which are consequently perceived as irrelevant and marginal[8].

In the end, the representation of semantic relations in ancient lexicons can differ significantly from modern conceptualizations.

My work on synsets is part of the PRIN 2022 "Resilient Septuagint". This project addresses the issue of intertextuality within the Biblical tradition, which is characterized by textual granularity and variety[9]. Its main goal is the development of a semantic search engine capable of managing this complexity and detecting Biblical citations within Greek texts. In this context, the semantic fields that we focus on are killing, healing, dream, and vision.

Synsets are crucial for capturing all relevant conceptual nuances, managing the granular and plural nature of Biblical texts—which never circulated in a single form—and understanding their relationship with other texts, such as early Christian commentaries.

2 Methodology

Inaccurate synsets can compromise NLP tasks, distorting model results and reducing coherence. When synsets are imprecise, NLP algorithms may misclassify or misinterpret textual data, resulting in erroneous outputs.

Consider a scenario where a lemma such as *apotheke* is mistakenly associated with a modern concept like "baseball" due to an erroneous synset definition. In this case, the model might misinterpret references to "storerooms" or "repositories" in ancient texts as related to contemporary concepts like "baseball". This mismatch could lead the model to inaccurately link unrelated passages or misrepresent the thematic content of the texts. For example, a passage discussing the storage of agricultural goods might be incorrectly connected to a text about sports.

An attempt to improve the precision of the Ancient Greek WordNet was made by Bizzoni et al. in 2015, focusing on a subset of lemmas based on the Homeric lexicon. This effort utilized distributional semantics to identify and correct lexical relationships through automated analysis of aligned corpora. [10]. However, manual refinement was also necessary to achieve the desired level of accuracy and reliability.

While a complete manual cleanup of the synsets—by assigning each lemma the most accurate definition—is theoretically possible, it is cost-effective due to time constraints. A practical approach is to begin by trimming down semfields, eliminating those that are completely out of context or anachronistic.

The elimination of the anachronistic categories was carried out in blocks of 100, based on the DDC, and evaluating the historical coherence of each with respect to the ancient Greek world. Once this step was completed, I conducted a general review, focusing on semfields that were similar or intersecting, and selecting the least misleading one.

This type of cleanup was performed across the entire database. To further refine the data regarding our fields of interest, I worked on a smaller group of lemmas selected by the philologists Laura Bigoni and Fabio Tutrone. For each lemma, which had previously been assigned several synsets of varying degrees of coherence by AGWN, I retained three synsets that best reflected the semantics of the lemmas, respecting their polysemy.

In this paper, I have focused mainly on the first part, relating to the analysis of semfields. Below, some examples of the semfields that could be removed from WordNet will be presented and discussed.

3 Anachronisms

Certainly, it is possible to identify and eliminate semfields that are clearly anachronistic. However, some semfields, while including modern and contemporary concepts, may still contain references to the ancient culture.

Semfields that can be safely eliminated are those related to modern and contemporary entities, institutions, and geopolitical concepts with no continuity with the ancient world, in order to mantain contextual and historical relevance. For example, many semantic fields in class 900 (Geography and

been updated periodically since then to reflect changes in knowledge and terminology; M. Dewey (1876), *Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library [Dewey Decimal Classification]*. Project Gutenberg, 2004. Retrieved July 15, 2024 from https://www.gutenberg.org/files/12513/12513-h/12513-h.htm

[8] Igwe - Ayandokun 2024, 214-227.

[9] Dainese - Mambelli 2024, 40-41.

[10] Bizzoni et al. in 2015, 47-50.

History) pertain to modern states such as those in North, South, and Central America, or Oceania. This includes categories related to American culture and the New World, such as 897 (Native North American literatures) and 817 (American humor and satire in English). The same applies to other modern languages and their literary works, including Spanish, French, Italian, and Russian.

Additionally, semfields describing modern and contemporary historical events tied to these geopolitical entities, such as 940.2 - covering the historical period from 1453 onward, the early modern period - should also be excluded[11].

Another macro group includes objects, tools, technologies, and artistic or scientific techniques that have no equivalent in Ancient Greece. The most evident examples pertain to the digital and virtual worlds, such as the semantic fields 770 (Photography, computer art, cinematography, videography), 384.5 (Wireless communication), and everything related to IT and computing, primarily found in class 000 (Computer science, information, and general works).

This group also includes concepts tied to modern or contemporary inventions, such as 384.6 (Telephone), 358.1 (Guided missile forces and warfare), and 686 (Printing and related activities), which presuppose the invention of telephone communication, guided missile warfare, and printing[12].

Other examples come from the world of music, with semfields related both to musical genres, such as 784.5 (Pop Music), and to musical instruments that have no counterpart in ancient culture, such as 786.7 (Electrophones, electronic instruments).

However, more generic categories might still encompass elements relevant to both the ancient and modern worlds, such as 784.1 (General principles, musical forms, instruments) or semfields related to specific types of instruments that already existed in antiquity, like string, wind, and percussion instruments.

This demonstrates that some semfields can indeed contain elements useful for analyzing ancient languages. For instance, music is a form of artistic and cultural expression documented since antiquity. A similar argument could be made for philosophy. Perhaps more than any other intellectual discipline, philosophy spans centuries, continually exploring themes and questions rooted in ancient thought. It would not be surprising, therefore, to connect the term *episteme* to the epistemological doctrines of Kant and Wittgenstein, just as *psyche* relates to the ideas of Freud and Jung[13].

In other cases, elements consistent with the Hellenic context can also be found in categories related to much more modern concepts. I have found that this holds true mostly for disciplines and sciences that have a modern status but address issues that were already subjects of inquiry in ancient times.

Examples of this type can also stem from the already mentioned term *psyche*, with its various diachronic nuances. Indeed, some of the most interesting entries relate to the semantic field of perception, movement and vision, or even to subfields concerning the subconscious and altered states of mind. These are all psychological aspects that Greek language can describe meticulously with its rich and polysemous lexicon.

Just consider the variety of verbs used to indicate actions related to sight. The verb *horaw* alone encompasses a wide range of visual experiences, from simple sight to perceiving something with understanding, while *skeptomai* is entrusted with indicating careful observation and examination, and *theaomai* carries a connotation of contemplation and admiration. In contrast, *blepo* refers to a more active and direct form of observation[14].

Clearly, not all the categories of this section can be applicable or useful in ancient studies. The semfield 152.8 Quantitative threshold, discrimination, reaction-time studies, for example, refers to modern studies involving specific concepts and methodologies, that have no ancient equivalent.

Another emblematic case is the section 610-619, which covers 'Medicine and Health.' Undoubtedly, semfields like 612 Human Physiology and 616 Diseases will provide fundamental lemmas such as *kardia, soma, nous* or *nosos*, *ousia* and *algos*.

Similarly, categories related to wounds and injuries, as well as surgical, pharmacological, and therapeutic techniques in general, are relevant. While these categories may include many references to modern technologies—such as MRI or contemporary surgical methods—that lack ancient counterparts, eliminating these semfields entirely

[11] 940.2 includes historical events up to 1814.

[12] Although specific technologies and weapons similar in terms of action and effect were employed, one cannot speak of "guided missile forces" in Ancient Greece; Partington 1999, XVI considers Greek fire and gunpowder represent «pre-modern forms of "scientific" knowledge».

[13] Wright - Potter 2000, 1-11; Brunschwig 2008, 229-240.

[14] Bran 2014, 216-222.

could result in the omission of important lemmas. These might include terms for body parts, types of wounds or cuts, various forms of trauma, and specific terminology related to military or scientific contexts, particularly those characteristic of Hippocratic medicine[15].

Whithout a doubt, it makes sense to handle semfields related to more general categories (e.g., 617 Surgery I& Related Medical Specialties) with greater caution and to disregard those that are more specific (e.g., 617.9 Geriatric, Pediatric, Military, Plastic Surgery, Transplantation of Tissue and Organs, Anesthesiology). Other examples may shed light on this approach.

In the case of engineering, which falls under sections 620-629, it is clear that we will find unequivocally anachronistic semfields related to advanced technologies, such as 629 Aerospace Engineering, or to modern concepts and inventions, like 621 Applied Physics and 626 Highway Engineering. In other cases, however, we will find semfields applicable to multiple historical periods, such as 623 Civil Engineering, or very general ones, such as 620.1 General Principles of Engineering.

If we take a semfield like 658.3 Personnel Management, the situation is different. We are dealing with a category that is already very specific and narrow in semantic and cognitive terms. Undoubtedly, the concept of personnel management - and even less so that of human resources - was not formalized as it is today. However, forms of it did indeed exist, from armies to state bureaucracy to public construction. It is certainly not difficult to imagine that within such a semfield, prominent lemmas like *hegemon, logistes* or *therapon* could be included, as they denote specific social roles.

Moreover, personnel and human resource management is a complex and interdisciplinary domain: it spans various disciplines, both ancient and modern, including sociology, philosophy, and economics. Moreover, many administrative and organizational aspects—although general in nature—are not in contradiction with the ancient world. This is true for processes such as selection and recruitment, for example of soldiers, or for training, such as that of future politicians[16].

Due to its less formalized nature compared to modern practices and its adherence to a different societal context, it is clear that this particular semfield will, in most cases, contain terms related to the modern world. Given that the semfield is already quite specific and narrowly defined, if the decision is not to eliminate it entirely, it would be advisable to manually clean the individual lemmas by associating them with the correct synsets.

On the other hand, it is quite plausible that some lemmas might appear in multiple semantic fields[17]. For example, to simplify, the lemma *psyche* could be relevant both to philosophical contexts and to modern psychological concepts.

This requires a careful evaluation of lemmas' applicability across different semantic fields. When a lemma appears in more than one field, it may be advantageous to assess which semantic field provides the most historically accurate and contextually appropriate representation, and to retain that one while disregarding the other. Given the large number of lemmas in AGWN, automating this process can be very useful.

4 Conclusion

Refining the Ancient Greek WordNet (AGWN) necessitates a careful balance between historical accuracy and semantic preservation. It is crucial to prioritize the removal of clearly anachronistic material. Simultaneously, retaining synsets and semfields that accurately reflect ancient contexts—despite potential overlaps with modern interpretations—is essential. This is particularly relevant for broader categories, which are more likely to contain diachronically valid and less specific elements.

Manual refinement is vital for detailed fields where anachronistic and historically accurate terms coexist; this is where biases can be most pronounced. Terms related to modern technologies or concepts without ancient counterparts must undergo rigorous review and be aligned with appropriate synsets[18]. This meticulous process minimizes the risk of misrepresentation and enhances the semantic accuracy of the database. Moreover, cross-checking problematic lemmas ensures that they are

[15] Until the last century, Greek medicine was often regarded as the foundation of modern biomedical science, celebrated for its rationality and clinical observations. Although this idealized view constrained a deeper understanding, it highlights a fundamental connection between ancient and modern medicine, demonstrating significant continuity and influence; van der Eijk 2005, XIV-XVI.

[16] Finley 1973, 17-34.

[17] By the way, a single lemma can be associated with multiple synsets; Biagetti et al. 2021, 259.

[18] Lemmas can be manually associated with other definitions, but typically WordNets, including AGWN, also provide the option to create new synsets and definition.

not confined solely to misleading semantic fields, thereby preserving valuable material.

The improvements in AGWN offer scholars more reliable tools for exploring ancient texts performing NLP tasks more accurately. They enable more nuanced insights into the semantic and intertextual relationships that have shaped the Greek literary traditions.

Moreover, this study highlights the theoretical importance of scrutinizing how modern frameworks can inadvertently impose contemporary biases on ancient knowledge. In doing so, it calls for a more critical use of digital resources in the study of antiquity, ensuring that the tools we rely on reflect the historical realities they aim to represent.

5 References

Luisa Bentivogli, Pamela Forner, Bernando Magnini, and Emanuele Pianta. 2014. Revising the Wordnet Domains Hierarchy: semantics, coverage and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Resources*, pages 94-101.

Erica Biagetti, Chiara Zanchi and William Michael Short. 2021. Toward the creation of WordNets for ancient Indo-European languages. In *Proceedings of the 11th Global WordNet Conference 2021*, pages 258–265.

Yuri Bizzoni, Federico Boschetti, Riccardo Del Gratta, Harry Diakoff, 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.

Yuri Bizzoni, Riccardo Del Gratta, Federico Boschetti, and Marianne Reboul. 2015. Enhancing the Accuracy of Ancient Greek WordNet by Multilingual Distributional Semantics. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it*, pages 47-50.

Răzvan Bran. 2014. From Sight to Thought. A Diachronic View On the Greek Verbs of Cognition. *Research and Science Today* 2(8):216-222.

Jacques Burschwig. 2008. The epistemological turn. In *The Cambridge History of Hellenistic Philosophy*, pages 29-41.

Davide Dainese and Anna Mambelli. 2024. Intertestualità tra Bibbie e antichi commentari cristiani: l'esempio di simul nel De Genesi ad litteram di Agostino. *Lexicon Philosophicum. Internationl Journal for the History of Texts and Ideas* 11:40-65.

Melvil Dewey. 2004. Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library [Dewey Decimal Classification], *The Project Gutenberg EBook*.

Philip van der Eijk. 2005. *Medicine and Philosophy in Antiquity*. Cambridge University Press, Cambridge, UK.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, MIT Press, Cambridge, Ma.

Christiane Fellbaum. 2005. WordNet and Wordnets. In *Encyclopedia of Language and Linguistics*, pages 665-670.

Kingsley N. Igwe and Ahmed Abayomi Ayandokun. 2024. Review of the treatment of religion and religious works in the library of congress and Dewey decimal classification schemes for knowledge organization in libraries. *Samaru Journal of Information Studies* 24:214-227.

Stefano Minozzi. 2009. The Latin WordNet Project. In *Proceedings of the Fourth International Global WordNet Conference. GWC 2008 (Szeged, Hungary, 22nd-25th January 2008)*, pages 707-716.

Moses Israel Finley. 1973. *The Ancient Economy*, University of California Press, Berkeley, California.

James Riddick Partington. 1999. *A History of Greek Fire and Gunpowder*, The John Hopkins University Press, Baltimore, Maryland.

John Paul Wright and Paul Potter. 2000. *Psyche and Soma: Physicians and Metaphysicians on the Mind-Body Problem from Antiquity to Enlightenment*, Oxford University Press, New York.