Evaluating Hierarchical Aggregation and LLM-Based Matching for Synset Selection in Ancient Greek

Luca Brigada Villa[1], Marco Passarotti[2], Chiara Zanchi[1], Riccardo Ginevra[2], Erica Fratellini[2], Eleonora Litta[2]
[1]University of Pavia
Piazza del Lino, 2, 27100 Pavia, Italy
{luca.brigadavilla, chiara.zanchi}@unipv.it
[2]Università Cattolica del Sacro Cuore
Largo A. Gemelli 1, 20123 Milan, Italy
{marco.passarotti, riccardo.ginevra, erica.fratellini, eleonoramaria.litta}@unicatt.it

Abstract

This paper presents a structured framework for WordNet synset selection applied to Ancient Greek lexical material. Starting from synonym definitions extracted from the Liddell–Scott–Jones (LSJ) lexicon, we compare two strategies: hierarchy-driven aggregation via bounded hypernym trees and LLM-based definitional matching with pairwise ranking. Graded human evaluation shows that structure-aware methods provide a robust baseline, particularly for nouns and verbs, while LLM-based reranking does not consistently improve performance, especially for highly ploysemous groups of synonyms. Beyond supporting the development of an Ancient Greek WordNet, the study highlights the methodological portability of the framework to other languages and lexical resources.

Keywords: WordNet, synset selection, Ancient Greek, LLMs

1. Introduction

Lexical-semantic resources such as WordNet (Miller, 1992; Fellbaum, 1998), FrameNet (Baker et al., 1998), and multilingual initiatives such as BabelNet (Navigli and Ponzetto, 2010) and Concepticon (List et al., 2025) provide structured inventories of senses organized into semantic networks or hierarchical taxonomies. These resources are widely used in linguistic research and natural language processing, where tasks often require linking lexical material to an appropriate sense. Despite the apparent simplicity of this objective, selecting a representative sense for a set of semantically related lexical items remains challenging, especially when the items reflect subtle sense distinctions or span multiple regions of the semantic hierarchy. The present work addresses this problem in the context of synonym groups derived from Ancient Greek lexical material, using definitions extracted from the *Liddell–Scott–Jones* (LSJ) lexicon as the basis for alignment to WordNet synsets.

A central difficulty arises from polysemy and semantic dispersion. Groups of near-synonymous words may correspond to closely related but distinct synsets, or they may cluster around different abstraction levels within the WordNet hierarchy. As a result, identifying a single synset that best captures the intended sense involves balancing structural coherence within the taxonomy and semantic similarity at the level of glosses.

In this work, we investigate two complementary strategies for synset selection. The former leverages the hierarchical organization of WordNet by constructing bounded hypernym trees and ranking candidate roots according to the portion of the retrieved semantic space they organize. The latter relies on semantic comparison between candidate glosses and a derived metadefinition, using large language models (LLMs) to perform pairwise preference judgments aggregated through an Elo ranking scheme.

We evaluate these strategies across multiple experimental conditions, including different parts of speech and both monosemous and polysemous synonym groups. Using graded human annotation and position-weighted scoring, we compare the effectiveness of structure-based and LLM-based approaches.

The paper is structured as follows. Section 2 situates the study within existing approaches to lexical-semantic alignment and synset selection. Section 3 presents the methodological framework, including the construction of synonym groups, definition extraction, Bag of Definitions (BoD) representation, synset retrieval, hypernym-tree structuring, and the two candidate selection strategies, together with the human evaluation protocol. Section 4 reports quantitative results, and Section 5 discusses their implications. Section 6 concludes and outlines directions for future research.

2. Related Work

Research on synset selection and lexical alignment has developed at the intersection of lexical semantics and natural language processing.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 368–379.

In linguistic theory, questions of sense granularity, polysemy, and semantic field structure have long shaped debates about how lexical meaning should be partitioned and organized (Geeraerts, 2001, 2007; Taylor, 2003). In computational settings, these issues surface in tasks such as word sense disambiguation, gloss matching, and taxonomy alignment, where the goal is to associate lexical material with entries in structured semantic inventories such as WordNet. Selecting an appropriate synset for a group of near-synonymous forms requires reconciling fine-grained lexicographic distinctions with higher-level taxonomic abstraction.

Within natural language processing, this problem has been studied primarily under the umbrella of word sense disambiguation and gloss-based matching. Early approaches relied on lexical overlap between context and glosses, most notably variants of the Lesk algorithm (Lesk, 1986; Banerjee and Pedersen, 2002). Subsequent work incorporated distributional semantics, enabling similarity comparison between contextual or definitional representations in vector space (Mikolov et al., 2013; Pennington et al., 2014). More recently, contextualized encoders have been employed to model interactions between glosses and lexical contexts (Huang et al., 2019), further refining sense selection mechanisms.

Beyond local similarity, several studies have emphasized the importance of taxonomic structure in guiding sense choice. The hierarchical organization of WordNet has been used to constrain disambiguation, promote semantic coherence, and support alignment across resources (Navigli, 2009; Ponzetto and Navigli, 2009). From a linguistic standpoint, hierarchical relations provide a principled means of balancing abstraction and specificity, though care must be taken to avoid systematic preference for overly general synsets.

Cross-lingual lexicoraphic alignment, especially when involving historical languages, introduces additional complexity. Mapping definitions from classical or non-English lexica to English-based resources such as WordNet requires mediating between distinct lexicographic traditions and potentially divergent sense inventories. Prior work on multilingual WordNets and lexical alignment (Bond and Foster, 2013) highlights both the potential and the challenges of projecting sense distinctions across languages.

More recently, large language models (LLMs) have been applied to definitional comparison and semantic matching tasks, including gloss-based word sense disambiguation and definition ranking. Work on contextualized sense embeddings has shown that transformer models can align lexical items with WordNet glosses, inducing sense-specific representations through similarity between distributional and definitional embeddings (Loureiro et al., 2022; Scarlini et al., 2020). In these approaches, synset selection is typically framed as direct similarity scoring or comparative ranking of glosses. While such methods effectively capture semantic relatedness at the level of definitional content, they generally abstract away from the explicit hierarchical organization of the lexical taxonomy. The interaction between hierarchical structure and LLM-driven semantic comparison remains comparatively underexplored.

The present study contributes to this line of research by explicitly contrasting structure-based aggregation within the WordNet hierarchy with LLM-based semantic matching, evaluated through graded human annotation in the context of Ancient Greek lexical material aligned to WordNet.

3. Methodology

Our methodology addresses the problem of mapping synonym groups to WordNet synsets by relying on definitional evidence rather than on lexical items alone. Starting from dictionary definitions associated with the lemmas in each group, we construct an intermediate representation that captures the semantic content of the group as a whole. This representation is then used to identify and organize candidate synsets from the target inventory. The following subsections describe the construction of the data representation, the procedures used to retrieve and structure candidate synsets, and the methods adopted to select and evaluate final candidates.

3.1. Synonym groups

We start from 16 synonym groups previously extracted by Marchesi et al. (2025)[1]. The dataset (Appendix A) is balanced by part of speech, comprising four groups for each of the following categories: verbs, nouns, adjectives, and adverbs. The 16 groups are further balanced with respect to lexical ambiguity, with eight groups populated by predominantly monosemous lemmas and eight by predominantly polysemous lemmas. Each group contains a set of Ancient Greek lemmas intended to be near-synonymous in at least one sense. Throughout the paper, the unit of analysis is the *group* rather than individual lemmas, since the final target is a synset-level mapping for the group as a whole. Groups are constructed based on shared definitional content rather than strict synonymy. They should therefore be interpreted as

[1]The dataset was created by extracting data from back-translation dictionaries and it is available at https://github.com/unipv-larl/llms-ag/tree/main/Data%20for%20fine%20tuning.

semantically coherent clusters reflecting overlapping regions of meaning, rather than sets of fully interchangeable lexical items. This perspective aligns with a relatively shallow notion of synonymy, whereby lexical items are grouped based on partial substitutability or shared semantic content rather than full equivalence, as is also the case in WordNet synsets (Fellbaum, 1998).

3.2. Definition extraction from LSJ

For each lemma in a synonym group, we extract all definitional material available in the corresponding entry of the online version of the LSJ (Liddell et al., 1996). LSJ entries are provided in a structured XML format, where definitions are organized into `senses` (and potentially nested `subsenses`). For each lemma, we collect the list of English definition fragments associated with every sense and subsense, without attempting prior lemma-level sense disambiguation. This choice is deliberate: LSJ definitions are often highly granular and fragmentary, and synonym groups frequently conflate multiple closely-related senses, making pre-disambiguation both costly and error-prone at this stage.

The output of this step is a multiset of short English definition strings per lemma. These are subsequently merged at the group level (Section 3.3).

3.3. Bag of Definitions representation

We represent each synonym group via a *Bag of Definitions* (BoD), defined as the multiset union of the LSJ definitions extracted for all lemmas in the group:

$$\text{BoD} = \{d_1, d_2, \ldots, d_n\}$$

where each d_i is an English LSJ definition. Intuitively, the BoD acts as a semantic fingerprint for the synonym group, aggregating definitional evidence across lemmas while remaining agnostic to within-lemma sense boundaries.

3.4. Synset retrieval via BoD–gloss matching

The mapping task is formulated as a retrieval problem: given a BoD query representing a synonym group, we rank all candidate synsets in Open English WordNet (OEWN) by semantic similarity between the BoD text and a synset textual representation.

Synset representations. For each OEWN synset, we build a short document consisting of its gloss. This choice aligns the target representation with the definitional nature of the BoD.

POS filtering. Since synonym groups are POS-homogeneous, we restrict retrieval to synsets with the corresponding POS (verb, noun, adjective, adverb). This reduces the search space and mitigates systematic errors due to cross-POS ambiguity.

Bi-encoder models. We compute BoD–synset similarity using four sentence embedding models:

- `all-mpnet-base-v2` (Song et al., 2020; Reimers and Gurevych, 2019),
- `e5-base-v2` (Wang et al., 2022),
- `bge-base-en-v1.5` (Xiao et al., 2023),
- `all-MiniLM-L6-v2` (Wang et al., 2020; Reimers and Gurevych, 2019).

Each model independently encodes the BoD query and the synset document into a fixed-dimensional vector; synsets are ranked by cosine similarity. For each group we obtain a complete ranking over the POS-filtered synset inventory.

3.5. Hypernym-tree structuring of top-ranked synsets

Retrieval produces a ranked list, but the highest-ranked region may still contain semantically heterogeneous candidates, especially when BoDs mix multiple related senses. To introduce structure, we compute hypernym trees for synsets in the top region of the ranking and use these trees as an abstraction mechanism.

All WordNet access and hierarchical traversals were implemented using the `wn` Python library (Goodman and Bond, 2021), which provides programmatic access to synsets and lexical relations.

Hypernym trees can be constructed only for nouns and verbs, since in WordNet these parts of speech are organized into well-formed hierarchical taxonomies. Adjectives and adverbs, by contrast, lack a comparable hypernym–hyponym structure in WordNet and are instead connected primarily through non-taxonomic relations (e.g. similar-to, pertains-to). For this reason, the procedure described in this section applies exclusively to nominal and verbal synsets.

For the purposes of this work, a hypernym tree is defined as the subgraph induced by taking a synset as the root and iteratively following the hyponym relation in the target inventory. Starting from an individual synset, its hypernym tree therefore consists of the synset itself and the set of more specific synsets that fall under it in the WordNet hierarchy, organized as a rooted structure in which lower nodes correspond to increasingly specific semantic categories. In this sense, hypernym trees

capture the semantic scope covered by a synset, rather than its ancestry, and provide a way to characterize how broadly or narrowly a candidate synset organizes the retrieved semantic space.

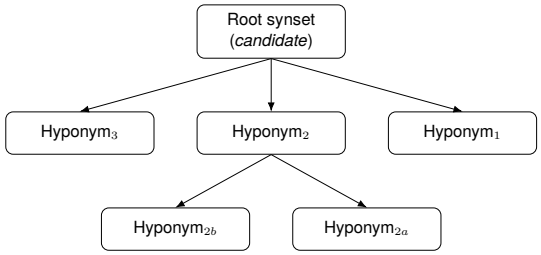

Figure 1: Schematic hypernym tree used as an abstraction device. A candidate synset is treated as a root, and the tree expands downward by iteratively following hyponym relations, capturing the semantic scope organized by the root.

By projecting retrieved synsets onto the hypernym trees rooted at candidate synsets, we move from a flat ranked list to a structured representation in which candidates can be compared in terms of the portion of the hierarchy they subsume. This abstraction makes it possible to reason about candidate synsets not only in terms of individual similarity scores, but also in terms of their capacity to organize related candidates within the lexical hierarchy. To avoid selecting candidates with excessively broad semantic scope, we restrict the procedure to hypernym trees rooted at synsets that have at least one hypernym; synsets without hypernyms are therefore excluded from consideration as roots. Concretely, given the top-100 synsets retrieved for a synonym group, we construct all possible three-level hypernym trees in which these candidates participate; that is, for each synset we consider (i) the tree rooted at the synset itself, (ii) the tree rooted at its direct hypernym, and (iii) the tree rooted at the hypernym of that hypernym. With reference to the schematic representation in Figure 1, synsets from the top-100 retrieved set may occupy different positions within the constructed trees: they can serve as roots, appear at intermediate levels, or occur among the lower-level hyponyms.

Because trees are constructed independently for each eligible root, partial structural overlap between trees may arise (e.g., when a synset and its hypernym both generate trees). We deliberately do not merge such overlapping subtrees into deeper composite structures. Merging would systematically favor higher-level synsets in the hierarchy, as they would inherit coverage from multiple lower-level trees and thus accumulate artificially inflated scope. By keeping trees distinct and limiting their depth, we ensure that each candidate is evaluated on the basis of how much of the top-100 retrieved set is locally organized within at most two hyponymic levels, rather than along arbitrarily long hypernym chains.

3.6. Candidate selection strategies

We compare two strategies for selecting final candidate synsets from the retrieved set. The applicability of these strategies depends on the part of speech, reflecting structural differences in the organization of WordNet. In particular, the hypernym-based strategy 3.6.1 applies only to nouns and verbs, for which hypernym–hyponym hierarchies are available.

3.6.1. Strategy A: hypernym-tree coverage

Strategy A is applied exclusively to nominal and verbal synsets. Rather than operating directly on the initially retrieved synsets, this strategy derives candidate synsets from the roots of the hypernym trees constructed over the retrieved set.

Starting from the top-ranked retrieved synsets, we construct a forest of eligible hypernym trees (cf. Section 3.5). Each distinct root of these trees – whether or not it appears in the original retrieved list – is treated as a candidate synset. Candidate synsets are ranked by the cardinality of the intersection between the original top-100 retrieved synsets and the set of nodes in their corresponding hypernym trees. Roots whose trees cover more retrieved synsets are preferred, as they provide broader yet structured abstraction over the retrieved semantic space.

3.6.2. Strategy B: LLM-based metadefinition matching

Strategy B selects and ranks candidate synsets by directly comparing their glosses to a metadefinition derived from the corresponding BoD, using pairwise judgments produced by LLMs.

Metadefinition construction. For each synonym group, the associated BoD is first transformed into a single metadefinition. This step is carried out by prompting three different LLMs, namely `Qwen2.5-3B-Instruct-GGUF` (Yang et al., 2024, 2025), `gemma-2b-it-GGUF` (Mesnard et al., 2024), and `Mistral-7B-Instruct-v0.2-GGUF` (Jiang et al., 2023) with identical instructions, in order to reduce model-specific biases and increase robustness (see Appendix B for the prompt text and parameters). The resulting metadefinitions provide an abstract semantic target against which candidate synsets are evaluated.

Candidate set. The set of candidate synsets depends on the part of speech. For nouns and verbs, candidates consist of all synsets that serve as roots of hypernym trees covering at least two synsets from the original top-100 cosine-similarity retrieval set (cf. Sections 3.4 and 3.5). This constraint ensures that only roots with a minimal degree of semantic support in the retrieved set are considered. For adjectives and adverbs, which lack a hypernym–hyponym hierarchy in WordNet, candidates are defined as the top-50 synsets ranked by cosine similarity.

Pairwise comparison. Candidate synsets are ranked via a sequence of pairwise comparisons. At each step, two candidates are sampled at random from the candidate set. The LLM is prompted with the metadefinition and the glosses of the two synsets, and asked to determine which gloss is semantically closer to the metadefinition (see Appendix C for the prompt text and parameters). The model is required to output a binary preference, without justification.

Elo-based ranking. The outcomes of the pairwise comparisons are aggregated using an Elo rating system (Elo, 1978). For each synonym group, the number of pairwise comparisons is set to five times the number of candidate synsets, resulting in a total of $5 \times |C|$ Elo matches, where $|C|$ denotes the size of the candidate set. Each candidate synset i is assigned an initial Elo score R_i, which is updated after each comparison. Given two candidates i and j, the expected score of i is computed as

$$E_i = \frac{1}{1 + 10^{(R_j - R_i)/400}}.$$

After observing the outcome of the comparison, the rating of i is updated according to

$$R'_i = R_i + K(S_i - E_i),$$

where $S_i \in \{0, 1\}$ indicates whether i is preferred by the LLM, and K is a scaling factor, set to $K = 24$ in our experiments. The rating of j is updated analogously. Over successive comparisons, synsets that are consistently preferred with respect to the metadefinition accrue higher Elo scores, resulting in a global ranking of candidates. An Elo-based ranking is preferred over direct similarity scoring because it aggregates relative judgments across many local comparisons, reducing sensitivity to score calibration and allowing a stable global ordering to emerge from noisy pairwise preferences.

3.7. Human annotation for evaluation

The outputs of all candidate selection strategies – hypernym-tree coverage for nouns and verbs (Strategy A), output of the bi-encoder models for adjectives and adverbs and LLM-based matching (Strategy B) – were evaluated through manual annotation.

For each embedding model and for each synonym group, we extracted the top-5 synsets produced by each strategy. Two independent annotators assessed the degree to which each proposed synset adheres to the semantic content of the corresponding synonym group. Annotations were assigned on a four-level ordinal scale reflecting increasing degrees of semantic relatedness: strong adherence (3), partial adherence (2), marginal adherence (1), and no adherence (0).

In addition, a third independent annotator selected one synset per synonym group that best represents the intended meaning of the group, without access to the experimental results. This independently chosen synset serves as a reference point for evaluating the alignment between the automatically ranked candidates and a human-identified representative sense.

4. Results

This section reports the empirical evaluation of the proposed strategies. We first assess inter-annotator agreement, then compare performance using both hit-based and graded score measures across embedding models, parts of speech, and monosemous versus polysemous groups. Given the limited size of the dataset (16 groups), results should be interpreted as indicative trends rather than as statistically significant comparisons.

4.1. Inter-annotator agreement

We assessed annotation reliability using Cohen's κ. Agreement was computed in two ways: (i) a binary version collapsing the scale into non-adherence (0) versus adherence (1–3), and (ii) a quadratic weighted κ over the full four-level ordinal scale (0–3), thereby accounting for the graded nature of the judgments.

Table 1 reports agreement per group. Across groups, binary κ yielded a macro-average of 0.506 (sd = 0.253), with values ranging from 0.109 to 0.820 and a pooled estimate of 0.520. Ordinal κ produced a slightly higher macro-average of 0.529 (sd = 0.311), with values ranging from 0.019 to 0.899 and a pooled estimate of 0.611.

Overall, agreement can be characterized as moderate, with substantial variability across groups. In general, monosemous groups tend to exhibit higher agreement than polysemous ones, suggesting that semantic dispersion within polysemous sets increases annotation difficulty. The higher pooled value observed for ordinal κ

		polysemous								monosemous							
	group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
κ	binary	0.654	0.259	0.707	0.205	0.622	0.534	0.182	0.354	0.820	0.790	0.762	0.724	0.146	0.790	0.109	0.434
	ordinal	0.655	0.269	0.889	0.249	0.694	0.770	0.082	0.198	0.744	0.899	0.776	0.810	0.158	0.867	0.019	0.377

Table 1: Inter-annotator agreement (Cohen's κ) per evaluation group. Binary κ refers to the 0 vs. 1–3 distinction; ordinal κ corresponds to quadratic weighted κ computed over the full 0–3 scale.

indicates that, even where exact level matches were inconsistent, annotators often agreed on the relative degree of adherence.

4.2. Hit-based evaluation

We first evaluate performance in terms of hit rate. A hit is recorded for a given embedding–strategy combination when at least one synset among the top-5 candidates is annotated by a human evaluator as showing at least marginal adherence (score ≥ 1) to the synonym group. For each combination, we report the proportion of groups for which a hit was observed. Hit rates are computed separately for each annotator and then averaged, ensuring that evaluation units remain independent at the group level.

Table 2 summarizes results across monosemous and polysemous groups as well as across parts of speech.

Overall, the non-LLM baseline ("none"), corresponding to hypernym-tree coverage for nouns and verbs and direct cosine top-5 selection for adjectives and adverbs, consistently achieves the highest hit rates across embeddings. In particular, the `bge-base-en-v1.5` model reaches near-ceiling performance for verbs and nouns, with hit rates of 1.00 in the two categories.

By contrast, LLM-based reranking generally reduces hit rates across embeddings. Among the three LLMs, Qwen tends to preserve performance more effectively than Mistral and Gemma, though results vary depending on embedding and part of speech. The drop in performance is particularly noticeable for polysemous groups, suggesting that reranking may introduce additional instability in semantically dispersed contexts.

Across parts of speech, nouns and verbs show higher hit rates overall, reflecting the structural advantage provided by hypernym-based aggregation. Adjectives and adverbs exhibit greater variability, especially under LLM reranking, likely due to the absence of hierarchical structure and stronger reliance on surface semantic matching.

To further assess alignment quality, Table 3 reports the proportion of groups for which the independently annotated gold synset – or one of its hypernyms (up to two levels up) or hyponyms (up to two levels down) – was retrieved among the top-5 candidates. This metric is substantially stricter, as it requires proximity to a specific reference synset rather than general semantic acceptability.

Under this criterion, performance decreases markedly across all embeddings and strategies. Even in the strongest configurations (e.g., `e5-base` without reranking), retrieval rates remain moderate and vary considerably across parts of speech. LLM-based reranking does not consistently improve gold retrieval and in several cases reduces it, particularly for polysemous groups. These results indicate that while multiple semantically acceptable synsets may be retrieved (as reflected in the hit-based evaluation), exact or near-exact alignment with a single gold synset is considerably more demanding.

Taken together, the hit-based and gold-based analyses suggest that structural aggregation supports the retrieval of semantically plausible candidates, but that convergence toward a specific target synset—especially in polysemous settings—remains challenging.

4.3. Graded score evaluation

While hit-based evaluation captures whether at least one acceptable synset is retrieved, it does not reflect how well relevant synsets are ranked within the top-5. We therefore compute a graded relevance score based on the four-level annotation scheme.

For each synonym group, each synset in the top-5 receives a relevance value in the range 0–3. Scores are aggregated using a normalized position-weighted scheme, assigning higher weight to higher-ranked positions and normalizing the resulting sum to ensure comparability across models and strategies.

Formally, let $r_k \in \{0, 1, 2, 3\}$ denote the relevance score assigned to the synset at rank k, for $k = 1, \ldots, K$ (with $K = 5$ in our setting). We adopt a logarithmic discounting scheme inspired by Discounted Cumulative Gain (DCG), assigning weight

$$w_k = \frac{1}{\log_2(k+1)}.$$

The normalized position-weighted score is then computed as:

$$S = \frac{\sum_{k=1}^{K} \frac{r_k}{\log_2(k+1)}}{\sum_{k=1}^{K} \frac{3}{\log_2(k+1)}}.$$

Embedding	+ LLM	poly	mono	verbs	nouns	adjs	advs
all-MiniLM-L6-v2	none	0.6875	0.8125	0.75	0.75	0.875	0.625
	+ mistral	0.4375	0.5625	0.5	0.625	0.375	0.5
	+ qwen	0.6875	0.625	0.75	0.875	0.625	0.375
	+ gemma	0.4375	0.5	0.5	0.625	0.375	0.375
all-mpnet-base-v2	none	0.6875	0.6875	0.625	0.5	0.875	0.75
	+ mistral	0.25	0.5	0.375	0.375	0.375	0.375
	+ qwen	0.5625	0.625	0.625	0.375	0.625	0.75
	+ gemma	0.3125	0.5	0.5	0.375	0.375	0.375
bge-base-en-v1.5	none	0.8125	0.875	1	1	0.625	0.75
	+ mistral	0.5	0.4375	0.625	0.375	0.375	0.5
	+ qwen	0.8125	0.6875	0.875	0.75	1	0.375
	+ gemma	0.5	0.4375	0.625	0.375	0.375	0.5
e5-base-v2	none	0.875	0.8125	0.75	1	0.75	0.875
	+ mistral	0.375	0.4375	0.25	0.625	0.375	0.375
	+ qwen	0.6875	0.625	0.75	0.625	0.625	0.625
	+ gemma	0.375	0.4375	0.25	0.625	0.375	0.375

Table 2: Proportion groups for which at least one valid synset was identified (per category and part of speech) across embedding models and reranking strategies. “None” denotes the absence of LLM reranking, corresponding either to the hypernym-based strategy (when applicable) or to direct top-5 selection by cosine similarity.

Embedding	+ LLM	poly	mono	verbs	nouns	adjs	advs
all-MiniLM-L6-v2	none	0.125	0.25	0.5	0	0	0.25
	+ mistral	0	0	0	0	0	0
	+ qwen	0.125	0.25	0.25	0.5	0	0
	+ gemma	0	0	0	0	0	0
all-mpnet-base-v2	none	0	0.25	0.25	0.25	0	0
	+ mistral	0	0.25	0	0.25	0	0.25
	+ qwen	0.125	0.125	0.25	0	0.25	0
	+ gemma	0	0.25	0	0.25	0	0.25
bge-base-en-v1.5	none	0.25	0.375	0.5	0.25	0	0.5
	+ mistral	0.25	0	0.25	0.25	0	0
	+ qwen	0.125	0	0	0	0.25	0
	+ gemma	0.25	0	0.25	0.25	0	0
e5-base-v2	none	0.25	0.5	0.5	0.5	0	0.5
	+ mistral	0.125	0.125	0	0.5	0	0
	+ qwen	0.25	0.5	0.5	0.25	0.25	0.5
	+ gemma	0.125	0.125	0	0.5	0	0

Table 3: Proportion of groups for which the gold synset – or one of its hypernyms (up to two levels up) or hyponyms (up to two levels down) – was retrieved among the top-5 candidates (polysemous vs. monosemous targets and by part of speech) across embedding models and reranking strategies. “None” denotes the absence of LLM reranking, corresponding either to the hypernym-based strategy (when applicable) or to direct top-5 selection by cosine similarity.

The denominator corresponds to the maximum attainable score (i.e., when all $r_k = 3$), ensuring that $S \in [0, 1]$. This formulation prioritizes correct ranking of highly relevant synsets at top positions, which is crucial in a manual selection scenario where only a small number of candidates is retained.

Table 4 reports normalized scores across embeddings, distinguishing between polysemous and monosemous groups as well as parts of speech.

Across all embeddings, the non-LLM baseline (“none”) consistently achieves the highest graded scores. This pattern mirrors the hit-based results but is even more pronounced, indicating that LLM-based reranking not only reduces the likelihood of retrieving at least one acceptable synset, but also tends to lower the overall concentration of semantically appropriate candidates at higher ranks.

Among reranking strategies, Qwen generally performs better than Mistral and Gemma, though it rarely surpasses the baseline. In some

Embedding	+ LLM	poly	mono	verbs	nouns	adjectives	adverbs
all-MiniLM-L6-v2	none	0.2018	0.1493	0.1699	0.1164	0.3010	0.114
	+ mistral	0.0394	0.0795	0.0510	0.0981	0.0513	0.0372
	+ qwen	0.1125	0.1378	0.0774	0.2697	0.1141	0.0395
	+ gemma	0.0424	0.0785	0.0510	0.0988	0.0513	0.0407
all-mpnet-base-v2	none	0.1464	0.1476	0.1115	0.1189	0.2865	0.0712
	+ mistral	0.0462	0.0779	0.0211	0.0994	0.0392	0.0885
	+ qwen	0.0552	0.1016	0.0925	0.0626	0.1037	0.0549
	+ gemma	0.0498	0.0779	0.0282	0.0994	0.0392	0.0885
bge-base-en-v1.5	none	0.2288	0.1226	0.1631	0.1746	0.1833	0.1818
	+ mistral	0.0598	0.0576	0.0688	0.0485	0.0267	0.0909
	+ qwen	0.1702	0.1111	0.1862	0.0943	0.2402	0.0419
	+ gemma	0.0598	0.0576	0.0634	0.0485	0.0321	0.0909
e5-base-v2	none	0.1674	0.1599	0.0931	0.1716	0.2157	0.1740
	+ mistral	0.0775	0.0752	0.0283	0.1344	0.0925	0.0503
	+ qwen	0.1182	0.1073	0.0556	0.1113	0.1442	0.1397
	+ gemma	0.0704	0.0697	0.0283	0.1233	0.0784	0.0503

Table 4: Normalized position-weighted scores (polysemous vs. monosemous and by part-of-speech) across embedding models and reranking strategies. "None" denotes the absence of LLM reranking, corresponding either to the hypernym-based strategy (when applicable) or to direct top-5 selection by cosine similarity.

cases, QWEN narrows the gap (e.g., nouns with `all-MiniLM-L6-v2` and adjectives with `bge-base-en-v1.5`), but improvements over the non-LLM strategy are limited and inconsistent.

With respect to monosemy and polysemy, performance differences are relatively small in the baseline condition, whereas reranking tends to amplify variability across groups. Polysemous sets show greater sensitivity to reranking, suggesting that pairwise LLM judgments may struggle when the retrieved candidates span multiple related semantic regions.

Part-of-speech differences remain visible in the graded scores. Nouns and verbs, which benefit from hypernym-based structural aggregation in the baseline condition, generally achieve more stable performance. Adjectives and adverbs exhibit greater fluctuation under reranking, consistent with their reliance on direct semantic comparison rather than hierarchical organization.

Overall, the graded evaluation confirms the robustness of the structurally grounded baseline and indicates that LLM-based pairwise reranking does not consistently improve ranking quality under the present configuration.

5. Discussion

The results highlight a consistent structural advantage for the hypernym-based strategy. Across embeddings and evaluation metrics, the non-LLM baseline – relying on hierarchical aggregation for nouns and verbs and direct cosine similarity for adjectives and adverbs – achieves the highest hit rates and graded scores. For nouns and verbs, these results suggest that organizing retrieved synsets through local hypernym structure provides a robust mechanism for concentrating semantically coherent candidates near the top of the ranking.

By contrast, LLM-based pairwise reranking does not consistently improve performance. Although pairwise judgments offer a flexible way to compare glosses against metadefinitions, they appear sensitive to semantic dispersion within the candidate set, particularly in polysemous groups. The Elo aggregation scheme stabilizes local comparisons, but the overall ranking remains dependent on the quality of individual LLM preferences, which may introduce additional variance.

Differences between monosemous and polysemous groups further reinforce this interpretation. Polysemous sets tend to fragment across multiple regions of the lexical hierarchy, reducing the effectiveness of both structural aggregation and semantic comparison. In such contexts, local hierarchical concentration appears more reliable than iterative pairwise reranking.

Overall, these findings suggest that explicitly leveraging lexical hierarchy remains a strong baseline for synset selection, while LLM-based semantic matching, at least in its current pairwise configuration, does not systematically outperform structure-aware methods. This points toward the potential benefit of hybrid approaches that combine hierarchical constraints with more globally cal-

ibrated semantic scoring.

6. Conclusion and Future Perspectives

This study compared hierarchy-driven aggregation and LLM-based definitional comparison for synset selection using Ancient Greek lexical material aligned to WordNet. The results indicate that structure-aware methods provide a strong and stable baseline, particularly for nouns and verbs, whereas LLM-based reranking does not consistently yield improvements, especially in polysemous contexts.

Beyond its specific contribution to the development of an Ancient Greek WordNet, the main strength of this work is methodological. The proposed framework – combining definitional extraction, hierarchical aggregation, and graded evaluation – is portable and applicable to other languages and lexical resources, offering a principled way to examine the interaction between synonym groupings and lexical hierarchy.

A limitation of the approach is its reliance on predefined synonym groups. Where curated synonym dictionaries are available, the method can be directly applied and may help reveal differences in semantic granularity across resources. In languages lacking such materials, synonym sets can be generated automatically with LLMs; although noisier, they provide a feasible starting point.

Future work will extend the framework cross-linguistically and explore hybrid pipelines in which automatically induced synonym sets are refined through hierarchical validation and human evaluation.

7. Bibliographical References

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 136–145. Springer, Berlin, Heidelberg.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Arpad E. Elo. 1978. *The Rating of Chessplayers, Past and Present*. Arco Publishing, New York.

Dirk Geeraerts. 2001. The definitional practice of dictionaries and the cognitive semantic conception of polysemy. *Lexicographica*, 17:6–21.

Dirk Geeraerts. 2007. Lexicography. In Dirk Geeraerts and Hubert Cuyckens, editors, *The Oxford Handbook of Cognitive Linguistics*, pages 1160–1175. Oxford University Press, New York.

Michael Wayne Goodman and Francis Bond. 2021. Intrinsically interlingual: The wn python library for wordnets. In *Proceedings of the 11th Global Wordnet Conference*, pages 100–107, University of South Africa (UNISA). Global Wordnet Association.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, page 24–26, New York, NY, USA. Association for Computing Machinery.

Daniel Loureiro, Alípio Mário Jorge, and Jose Camacho-Collados. 2022. Lmms reloaded: Transformer-based sense embeddings for disambiguation and beyond. *Artificial Intelligence*, 305:103661.

Beatrice Marchesi, Annachiara Clementelli, Andrea Maurizio Mammarella, Silvia Zampetta, Erica Biagetti, Luca Brigada Villa, Virginia Mastellari, Riccardo Ginevra, Claudia Roberta Combei, and Chiara Zanchi. 2025. Towards the Semi-Automated Population of the Ancient Greek WordNet. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 647–658, Cagliari, Italy. CEUR Workshop Proceedings.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak,

Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Simone Paolo Ponzetto and Roberto Navigli. 2009. Large-scale taxonomy mapping for restructuring and integrating wikipedia. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, IJCAI'09, page 2083–2088, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3528–3539, Online. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

John R. Taylor. 2003. *Linguistic Categorization*, 3 edition. Oxford University Press, Oxford.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei

Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

8. Language Resource References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Henry George Liddell, Robert Scott, Henry Stuart Jones, and Roderick McKenzie. 1996. A greek-english lexicon. `http://www.perseus.tufts.edu`. Digital edition, Perseus Digital Library.

Johann Mattis List, Annika Tjuka, Frederic Blum, Alžběta Kučerová, Carlos Barrientos Ugarte, Christoph Rzymski, Simon Greenhill, and Robert Forkel, editors. 2025. *CLLD Concepticon 3.4.0*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

A. Groups of synonyms

Group 1: ἀποκείρω, ἐπικρύπτω, κατέχω, καταπαύω, καταλύω, διαλύω, παύω, καταστρέφω, κλέπτω, κρύπτω, κατεργάζομαι, ὑποστέλλω.
Semantic domain: concealment, suppression, removal, end

Group 2: τρύχω, βασανίζω, στρεβλόω, παρακρούω, σπάω, ἑλκόω, ἀρταμέω, κνάπτω, διασπάω, διασπαράσσω, σπαράσσω, δάκνω, λυπέω, ἀλγύνω.
Semantic domain: physical or psychological affliction, damage

Group 3: τύπος, περιγραφή, χαρακτήρ, χάραγμα, ὑπογραφή, στίβος.
Semantic domain: mark, representation, inscription

Group 4: αἰθήρ, πόλος, ὄροφος, ὀροφή, οὐρανός, κέραμος, εὐδία, τέγος, ἀήρ, ῥυμός, Διοπετής, εὐφημία.
Semantic domain: upper space, sky, atmosphere, roof

Group 5: ἀλίγκιος, ἐμφερής, παραπλήσιος, ἴσος, προσφερής, ὅμοιος, προσόμοιος.
Semantic domain: similarity, equivalence, likeness

Group 6: χθόνιος, ὑπόνομος, κατάγειος, νέρτερος, ὕπουλος
Semantic domain: subterranean, hidden, lower realm

Group 7: πόθεν, πῃ, ποῖ, τάχα, ἄν.
Semantic domain: indefiniteness, uncertainty

Group 8: τέως, νῦν, τρίς, ἀεί, ἄλλοτε, ἐνίοτε, πηνίκα, καίριος, πολλάκις, ὁπηνίκα.
Semantic domain: (indefinite) temporal reference

Group 9: ῥυπάω, αὐχμέω.
Semantic domain: filth, neglect, squalor

Group 10: προοιμιάζομαι, ὑπαγορεύω.
Semantic domain: discourse initiation

Group 11: εὐθηνία, εὐπραγία.
Semantic domain: prosperity, success, well-being

Group 12: ὑποθήκη, φραδή, ὑπόνοια.
Semantic domain: suggestion, advice

Group 13: ἄκοπος, ἀκάματος.
Semantic domain: absence of fatigue, tirelessness

Group 14: ἀνήλιος, ἀναύγητος.
Semantic domain: absence of sun, darkness

Group 15: κάτωθεν, ἐκεῖσε, ἐκεῖ.
Semantic domain: direction, location

Group 16: ἑτέρωθι, ἀλλαχοῦ, ἄλλοθι, ἄλλος, ἑτέρωθεν, ἑτέρωσε, ἀλλαχόθεν.
Semantic domain: elsewhere

B. Prompt to generate the metadefinition from the BoD

Prompt:
```
You are a lexicographer.

Write exactly ONE definition (ONE
sentence) for the dominant sense in
the BoD.
Requirements:
- Start immediately with the
definition.
- No lists, no repetition, no
semicolons.
- KEEP IT SHORT!
- Max 25 words.
- Use ONLY information from the BoD.
- JUST ONE SENTENCE.
- STOP after the definition.

BoD:
{bod_block}

Definition:
```

Parameters: `max_tokens=80, temperature=0.2, top_p=0.9, repeat_penalty=1.15`

C. Prompt to select between two candidates

Prompt:
```
Metadefinition:
{metadef}

Gloss A:
{gloss_a}

Gloss B:
{gloss_b}

Which gloss matches the
metadefinition better?
Answer ONLY with ``A'' or ``B''.
```

Parameters: `max_tokens=2, temperature=0.0`