Towards a Gold Standard for Adjectival Hypernymy: Enriching the Open English WordNet with a Hybrid Approach

Lorenzo Augello[1], John P. McCrae[2], Marco Passarotti[1]
[1]Università Cattolica del Sacro Cuore, Milan, Italy
[2]Research Ireland Insight Centre and ADAPT Centre, University of Galway, Ireland
lorenzo.augello01@icatt.it, john@mccr.ae, marco.passarotti@unicatt.it

Abstract

Adjectival hypernymy is an underexplored lexical-semantic relation essential for Natural Language Processing (NLP) and hierarchical semantic organization of the lexicon. While hypernymy in nouns and verbs has been extensively modeled in resources such as WordNet, adjectives remain largely unstructured due to their gradability and context-dependence. We present a hybrid Large Language Model (LLM)-Human approach towards the creation of a gold-standard dataset for adjectival hypernymy. Our method integrates three LLMs with systematic human evaluation, guided by a specifically developed theoretical framework ensuring consistency and linguistically-based principles, compiling a resource of 3,836 validated adjective hyponym-hypernym pairs. Results demonstrate high precision for consensus predictions (87%), confirming the utility of cross-model agreement as a proxy for semantic validity. This method highlights how LLMs can complement human effort and expertise to support the construction of lexical resources. The resulting dataset aims to enrich the Open English WordNet (OEWN) with explicit adjectival hierarchies and serves as a benchmark for hypernymy detection and lexical entailment evaluation.

Keywords: Adjectival Hypernymy, Large Language Model Evaluation, Lexical Resource Enrichment, WordNet, English

1. Introduction

Lexical-semantic relations are fundamental for NLP and computational linguistics, supporting applications such as semantic search, question answering, information retrieval, and textual entailment. Among these relations, hypernymy - the "is-a" relation - has been the focus of extensive research, especially in the case of nouns and verbs (Hearst, 1992), constituting the hierarchical backbone of large-scale lexical resources such as WordNet (Fellbaum, 1998), enabling inference and knowledge organization among concepts.

Despite this crucial role in taxonomical reasoning, the systematic encoding of adjectival hypernymy remains largely underdeveloped. Adjectives differ from nouns and verbs in that their meaning is often gradable and context-dependent. For instance, adjectives such as *hard* or *cold* can vary in interpretation, and may belong to multiple conceptual dimensions (i.e., semantic domains): *hard* could be used in various contexts to define "a *difficult* problem" as well as "a *resistant* material", while *cold* may refer to both "a *rigid* weather" and "an *unfriendly* person".

These properties make it difficult to define clear hierarchical relations, which partly explains why major lexical resources - including the OEWN (McCrae et al., 2019) - do not yet provide explicit hypernymy links for adjectives. Nonetheless, encoding these relations is important both to improve the connectivity of lexical resources and to provide more thorough coverage of adjectival semantics.

Previous work on this (Augello and McCrae, 2025) led to the creation of a small initial gold-standard dataset of 302 adjective hyponym-hypernym pairs in English, derived from multilingual wordnets and validated through manual human annotation. However, the dataset remains limited in scope, preventing both systematic evaluation of NLP models and the broader integration of adjectival hypernymy into the OEWN. To make meaningful progress, it is essential to construct a larger thoroughly validated gold standard that can serve as both a benchmark for hypernymy discovery and generation, and a lexical resource enrichment.

With this work, we aim to create an evaluation resource for hypernymy detection that integrates the information already present in other lexical entailment benchmarks such as HyperLex (Vulić et al., 2017), which focuses only on nouns and verbs. Our purpose is to fill this gap in order to be able to evaluate systems on adjectival semantic relations and to propose a novel theoretical understanding of how adjectival category membership could be defined both in human cognition and NLP models.

The present study addresses these gaps through a hybrid approach that combines LLMs with human evaluation. We provide three models with input hyponym adjectives and their definitions, and we prompt them to generate candidate hypernyms, which are then manually validated following clear-cut theoretical guidelines. This enables us to scale up coverage while maintaining quality, and to assess the semantic competence of LLMs in handling complex lexical relations.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3662–3671
11–16 May 2026.

Our primary contributions are the following:

1. Definition of theoretical guidelines for adjectival hypernymy: we propose clear operational rules to address the fuzziness of this relation and ensure consistency in annotation, providing both a foundation for prompting LLMs and a reliable framework for human annotators, enhancing reproducibility and explainability.
2. Creation of an expanded gold-standard benchmark through a hybrid LLM-human approach, serving the two purposes of providing an evaluation set for automatic systems, and of adding adjectival data in current language resources and evaluation pipelines.
3. We categorise LLMs' outputs in three classes based on models' agreement (consensus, majority, singleton), and we show how the consensus method (cases where all models generate the same candidate hypernyms) gives a very high acceptance rate (87%), highlighting its reliability for automatic hypernymy generation.
4. Future enrichment of the OEWN, with the gold standard designed as an evolving resource that can be integrated into it to introduce explicit adjectival hypernymy links, improving its connectivity and coverage.

Focusing on the enrichment and curation of lexical resources, the development of reliable annotation guidelines, and the construction of benchmarks for evaluating computational systems, we aim at addressing the issue that lexical resources remain incomplete and in constant need of extension: our hybrid approach focuses on how emerging technologies can be used to accelerate resource construction and enhancement while preserving human control and quality, displaying scalable yet reliable enrichment. Ultimately, this study contributes both a concrete resource[1] and a methodological proposal for how LLMs and human meta-linguistic expertise can be combined to tackle difficult semantic relations.

The rest of the paper is structured as follows: Section 2 presents related work on both linguistic and theoretical issues, and empirical tasks and evaluation strategies. In Section 3 we present our theoretical framework for a clear definition of adjectival hypernymy, while Section 4 describes the methodology we used for the construction of the benchmark dataset. Its evaluation and validation are detailed in Section 5 and the final results are commented in Section 6. After summarising and concluding in Section 7, we present limitations and suggestions for future work in Section 8.

[1] https://github.com/lorenzoaugello/hypernymy

2. Related Work

2.1. Hypernymy and Adjective Semantics

Hypernymy is a hierarchical semantic relation in which a hypernym denotes a broader, more general category than its hyponym (Cruse, 1986; Murphy, 2003). This relation is clearly instantiated in nouns and their clear classes of semantic domains, and has been extensively represented in lexical resources such as Princeton WordNet (Miller, 1995; Fellbaum, 1998), where nouns and verbs' synsets are organized into structured hierarchies. Hypernymy has also been used as a core relation in semantic-related tasks such as knowledge extraction from the web, taxonomy induction, and textual entailment (Navigli et al., 2010; Vulić et al., 2017).

However, the extension of hypernymy to adjectives remains underexplored. Adjectives differ from nouns in several key aspects, as they often express scalar properties, are used to modify nouns rather than denote entities (Raskin and Nirenburg, 1995; Lyons, 1977), and show high contextual variability (Heyvaert, 2010). Their gradable and context-sensitive nature makes a classification within rigid taxonomic hierarchies difficult (Scheible and Schulte im Walde, 2014). As the behaviour of adjectives eludes a straightforward ontological modelling (McCrae et al., 2014), Princeton WordNet does not encode hypernymy relations for adjectives, which are rather linked by relations of similarity (Similar To and See Also relations) and antonymy.

In order to create taxonomies according to agreed criteria, early classification efforts (Dixon, 1982; Hundsnurscher and Splett, 1982) attempted to group adjectives into broad "supersenses" related to generic semantic classes (Tsvetkov et al., 2014), but these proposals remain at a high level of categorization. Nonetheless, research in multilingual lexical semantics has proposed operational solutions to represent adjectival hypernymy in lexical resources: wordnets for Polish (Maziarz et al., 2016), Dutch (Postma et al., 2016), and German (Hamp and Feldweg, 1997) have successfully organised adjectives into hierarchies, where broader attributes subsume more specific ones. These initiatives provide a basis for investigating adjectival hypernymy computationally, and Augello and McCrae (2025) follow this multilingual tradition by adapting existing hierarchies from Dutch and Polish to English and validating a first gold standard set of adjective pairs.

2.2. Lexical Entailment in Large Language Models

In the domain of lexical inference, hypernymy was initially approached through lexico-syntactic pattern

extraction (Hearst, 1992), but more recent studies have investigated whether and how LLMs encode such relations (Moskvoretskii et al., 2024). Ushio et al. (2021) show that transformer-based models like BERT can capture hypernymy structure, though often in inconsistent ways, lacking systematicity, especially beyond prototypical cases (Ravichander et al., 2020; Levy et al., 2015). Bouraoui et al. (2019) examine whether relational knowledge can be induced from BERT through prompting and fine-tuning, and Vulić and Mrkšić (2018) specialize word vectors for lexical entailment tasks, showing that targeted adaptation can improve semantic distinction.

However, these studies focus almost exclusively on noun-based relations, as well as several evaluation tasks (Bordea et al., 2016; Camacho-Collados et al., 2018). Beyond hypernymy, general lexical inference (i.e., deciding whether or not an entailment relation holds between two lexical items) has also been evaluated using contextualized tasks, showing that LLMs often struggle to distinguish closely related senses, especially in the case of polysemy in lexical semantics equivalence tasks (Hayashi, 2025). Similarly, Schmitt and Schütze (2021) argue that many models rely more on distributional similarity than on semantic entailment, failing to model directionality or hierarchical meaning.

2.3. Semantic Relations Benchmarks

The evaluation of semantic relations and lexical inference systems has traditionally relied on carefully designed benchmarks, which have taken multiple forms, from binary relation datasets to graded ones that capture finer semantic nuances.

One of the most influential early benchmarks is BLESS (Baroni and Lenci, 2011), focused on distributional semantics evaluation, testing both the identification of semantically related words and the specific semantic relation holding between them. This inspired a series of follow-up works such as EVALution (Santus et al., 2015), which extends coverage and includes many relation types such as antonymy, hypernymy, meronymy and synonymy, and ROOT9 (Santus et al., 2016), focused on distinguishing hypernyms from co-hyponyms and other *random* relations. Indeed, ROOT9 includes adjectives as well, so it is the closest method to the one adopted in this paper, but it is not specifically tailored on them (the parts of speech distribution is skewed towards nouns, as adjectives represent only 9% of the total hypernymy pairs) and does not provide clear theoretical guidelines, differently from one of our main purposes (see Section 3).

Other evaluation frameworks have incorporated graded semantic judgments, reflecting insights from cognitive psychology that semantic category membership is often a matter of degree. The most prominent example is HyperLex (Vulić et al., 2017), highlighting the discrepancy between human intuitions of semantic typicality and the binary predictions of many computational models. Another gold-standard resource for semantic evaluation is provided by SimLex-999 (Hill et al., 2015), which specifically targets the ability of models to reflect similarity.

Despite the high value of those resources, existing benchmarks almost exclusively focus on nouns and verbs, leaving adjectival semantics uncovered. The hyponym-hypernym correspondence is always *one-to-one*, limiting semantic richness, and not linguistically motivated. This omission is critical, as adjectives pose unique challenges due to their gradability, multidimensionality, and context-dependence. As a consequence, models and lexical resources lack systematic evaluation material for adjectival hypernymy, limiting both theoretical and practical progress.

3. Guidelines Definition

At the foundation of the dataset construction, defining a clear schema to describe adjectival hypernymy is necessary, highlighting its uniqueness and the need for a distinction from other semantic relations, grounded on specifically thought principles, and accounting for hierarchical reasoning, disambiguation, and part-of-speech (PoS) confusion, all tailored towards and enhanced by the OEWN.

We build on the initial guidelines proposed by Augello and McCrae (2025), which focus on the principle of substitution, the principle of inclusion of meaning, PoS ambiguity and polysemy. We expand them in order to create a clear framework useful for describing the semantic relation of hypernymy for adjectives, that could be helpful for the construction of lexical resources and a finer-grained hierarchical organization of lexical entities. In addition to a theoretical analysis and an operational aid to lexical resource development, we propose the following guidelines as a method for the evaluation of the linguistic capabilities of automatic systems in detecting taxonomic and semantic relations between adjectives:

- The hyponym and the hypernym must be different.
- Adjectives that are in the same synset need to have the same hypernym. Many adjectives are very similar and pertain to the same OEWN synset (e.g., *Eurasian*, *Eurasiatic*, oewn-03035646-a, "relating to, or coming from, Europe and Asia")[2] and must have the same hypernym.

[2] https://en-word.net/lemma/Eurasian

- Principle of substitution: if you substitute the hyponym with the hypernym, the meaning of the phrase should not change much. There could necessarily be some loss of specificity, but the difference should only concern a broader meaning. Vice-versa, if you substitute the hypernym with the hyponym, there could be loss of meaning because the scope of the hyponym is littler that the hypernym's one.
- Principle of inclusion of meaning: the meaning of the hyponym is narrower and should be included in the meaning of the hypernym, which is broader. Vice-versa, the meaning of the hypernym is not completely included in the hyponym's one.
- A hyponym and its related hypernym must not pertain to the same synset in OEWN, as this would entail a synonymy relation between them, rather than hypernymy.
- A hyponym could have more than one hypernym.
- Both the hyponym and the hypernym must be adjectives.
- If a hyponym lemma could have multiple PoS (e.g., see *clean* as adjective, adverb, noun and verb),[3] always consider the adjective one.
- For adjectives that are polysemous (e.g., *cold* as temperature-related or psychology-related),[4] always consider OEWN definitions to discriminate.
- Consider the following example for a distinction between synonymy and hypernymy in the OEWN: *happy* and *felicitous* are synonyms (under the definition, "marked by good fortune")[5] and can be substituted, e.g., "a happy life"/"a felicitous outcome". This does not mean that they can be substituted in every sense, e.g., "happy to help" but not "felicitous to help". This is valid for synonyms, but the substitution check must always hold for hypernymy.[6]
- Well-defined principle: it should be possible to easily write a definition for a lemma that is distinct from other lemmas in OEWN.

4. Methodology

4.1. Data Collection and Pre-processing

In order to compile an initial dataset of adjectives to gather the base input hyponyms for LLMs to generate hypernyms, two complementary lexical resources were chosen: the Kilgarriff list (Kilgarriff, 1997) from the British National Corpus (BNC) (Consortium, 2007) and the Core WordNet.[7] The Kilgarriff list is a lemmatized frequency list covering the 6,318 most frequent words occurring more than 800 times in the BNC. From this resource we extracted 1,124 adjective lemmas. The Core WordNet represents a set of concepts that constitute the most fundamental lexicalized ideas in WordNet, shared across multiple languages. From this, we extracted 698 adjective lemmas. Since both resources largely target the most frequent items in English, significant overlap was expected and, to avoid redundancy, we removed duplicates and also excluded all items that were already included in the 302-pair gold standard provided by Augello and McCrae (2025). The resulting dataset contained 1,148 adjective lemmas, ensuring broad lexical coverage while avoiding overlap with existing evaluation material.

An interesting challenge of the Core WordNet is that, for polysemous words, it contains duplicated lemmas corresponding to their distinct senses. For instance, the adjective *warm* corresponds to two entries, one associated to the definition "producing a comfortable degree of heat" and another to "friendly and responsive". Rather than collapsing these cases, we preserved all occurrences, allowing to address the problem of word-sense disambiguation in adjectival hypernymy. To increase semantic grounding and ensure that each word was associated with its correct meaning, we enriched each lemma with its synset definition extracted from the OEWN. For adjectives with a single occurrence, the definition of the first synset in the OEWN was selected automatically. For polysemous lemmas, we performed a manual alignment between the senses reported in the Core WordNet and the corresponding OEWN definitions, distinguishing sense-specific entries in the dataset. This allowed for providing the models with a grounded and accurate connection between words and their senses. The final outcome is a dataset of 1,148 adjective hyponyms with disambiguated definitions, ready to be used as prompts for LLMs in the hypernymy generation experiments.

4.2. LLMs and Prompting Strategy

For the hypernymy generation task, we selected three open-access LLMs: gemma-3:27b,[8] gpt-

[3] https://en-word.net/lemma/clean
[4] https://en-word.net/lemma/cold
[5] https://en-word.net/lemma/happy
[6] https://github.com/globalwordnet/english-wordnet/blob/main/NEW_SYNSETS.md
[7] https://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt
[8] https://huggingface.co/google/gemma-3-27b-it

Category	Generated pairs	Synset filtering	Count filtering	Final accepted	Acceptance rate
consensus	2,711	2,540	2,367	2,367	87%
majority	4,676	4,293	3,336	1,381	29%
singleton	14,634	14,076	2,433	88	0.6%
Total	22,021	20,909	8,136	3,836	17%

Table 1: Description of the size of the dataset, divided into each category set (consensus, majority and singleton), from the initially generated hypernymy pairs to the final kept ones, after filtering by synset and total count, annotating and thresholding.

oss:20b,[9] and mistral-small:24b.[10] The choice of the models was mostly driven by the principles of balance between size (kept below 30B parameters for computational feasibility) and performance, high linguistic capability, and availability of transparent evaluation.

The models were prompted to generate a list of adjective hypernyms (maximum 10) for each adjective hyponym in input. The definition for the hyponym was also explicitly included in the prompt, to reduce ambiguity and mitigate problems of polysemy. The theoretical guidelines reported in Section 3 were listed within the prompt, in order to guide the models towards outputs based on our theoretical criteria and as linguistically grounded as possible. The following is a prompt template:

> Given the hyponym adjective "{hyponym}" with definition "{definition}", generate a list of related adjective hypernyms. Only a list of adjective hypernyms must be in the output, nothing else more. Do not re-generate the input hyponym. Respect the following guidelines: [...]

To address the issue of stochasticity in LLMs outputs, we adopted a cross-fold prompting strategy. For each model, we ran four separate folds, re-prompting the same set of 1,148 hyponyms. This led to a total of 12 runs (3 models per 4 folds each), allowing us to assess stability across multiple generations.

4.3. Output Aggregation and Categorisation

In the cross-fold validation, the outputs of the four folds were aggregated for each model, resulting in three consolidated lists, one for each model, keeping track of the number of folds in which a candidate hypernym appeared. Combining different models through cross-fold replication ensured our evaluation not to be dependent on single-generation noise, but instead reflecting more stable patterns of hypernymy recognition. After the aggregation, we performed a cross-model comparison to identify convergences and divergences in the candidate outputs. We then classified the results into three distinct categories:

- Consensus set (2,711 pairs): candidate hypernyms generated at least once by all three models.
- Majority set (4,676 pairs): candidate hypernyms generated at least once by two out of three models.
- Singleton set (14,634 pairs): candidate hypernyms generated at least once by only one model.

Starting from this, as a first threshold, we removed all the pairs of adjectives that pertained to the same synset in the OEWN, but yet were incorrectly generated, deviating from one of the principles stated in the guidelines (see Section 3). This reduced the size of the three sets respectively to 2,540, 4,293 and 14,076 pairs. Those numbers were the starting point for our process of annotation and evaluation, which put together frequency-based principles and human validation, in order not just to blindly trust the models outputs, but to accept or reject them with a thorough approach designed for each different setting, leading to the final outcomes reported in Table 1.

5. Annotation and Evaluation

In order to define specific thresholds for validation and have an evaluation of the results principled on outputs statistics, for each candidate hypernym we recorded information about the number and name of the models that generated it, the number of folds in which it was generated, the number of times in which it was confirmed by each model, and its best, worse and average rank in the output list (see examples in Table 2). This information contributed to develop a set-specific evaluation methodology.

For the consensus set, we kept only the hypernyms that were confirmed at least once (so, generated in at least two folds) by each model, getting to a final number of 2,367 pairs that were accepted in the final dataset.

[9] https://huggingface.co/openai/gpt-oss-20b

[10] https://huggingface.co/mistralai/Mistral-Small-24B-Base-2501

Hypo	Hyper	Category	Models	Count	Folds	Avg rank	Best rank	Worst rank
ancient	*old*	consensus	gemma, gpt, mistral	11	4	1.25	1	2
abnormal	*uncommon*	majority	gpt, mistral	5	3	1.8	1	4
remote	*far*	singleton	gpt	4	4	2	2	2

Table 2: Examples of hyponym-hypernym candidate pairs with statistical infomation recorded across cross-fold validation, indicating the category set, the models that generated it, the total number of attestations, the total number of folds in which it was generated by each model, the average rank in lists across folds, the best and worst rank.

For the majority set, at a first stage we kept only the candidate hypernyms that were confirmed at least once by two models, getting to 3,336 pairs (71.3% of the initial total). We then performed manual annotation of a randomly selected 100-pair sample, with two expert annotators and a two-label reject/accept classification. With an Inter Annotator Agreement (IAA) of Cohen's k=0.89, 35 pairs were kept. For those, we analyzed the relation between frequency of generation (maximum of 8, 2 models per 4 folds each) and ranking stability (maximum of 10) and the probability of being accepted by human annotators, in order to set thresholds to validate the rest of the majority set.

Accepted items showed a significantly higher average count across folds (6.3) and lower average rank (3.4) than rejected ones (respectively 5.2 and 4.8). This difference is furtherly confirmed by the figures related to the best rank and worst rank average, respectively 1.9 and 5.2 for accepted candidates and 3.4 and 7.3 for rejected ones. This underlines once more the importance of human intervention in such a linguistically nuanced task, and shows how LLMs predictions with lower scores in the selected statistical indicators are in fact more likely to be refused by humans, indicating encouraging convergence between humans and systems' semantic interpretation.

Hence, building on this, we used these empirical distributions to define the two following rule-based thresholds for automatic acceptance: (i) Count ≥ 5 (candidates generated at least five times) and (ii) Avg_Rank ≤ 4 (candidates with an average rank not exceeding 4). Those two measures align with each other across outputs and are not mutually exclusive, as the higher is the total count, the lower is the rank. This configuration achieved a precision of 0.84 and recall of 0.72 on the annotated sample, providing a balance between reliability and coverage. The rule was then applied to the entire majority set to derive the final set of 1,381 automatically validated hypernymy pairs.

For the singleton set, we performed a further specifically-tailored evaluation, given the fact that this is necessarily the most uncertain setting. After taking out all the outputs generated less than 4 times (so, keeping only the ones confirmed by the single model in each fold) and getting to 2,433 pairs (16.6% of the original total), we performed a reverse hyponym generation.

Among the already tested models, we picked gpt-oss:20b, being it the model with the lowest number of singleton outputs (5%, compared to 8% of mistral-small:24b and 87% of gemma-3:27b), and we prompted it with the reversed task of hyponym generation. Given an input hypernym, if the generated hyponym was the same as the original one contained in the singleton set, then it was accepted, otherwise it was taken out. After reverse generation, only 88 pairs were confirmed (7.4% of the total 1,181 pairs),[11] corroborated by human evaluation performed by the same two annotators with an IAA of Cohen's k=0.73.

6. Results and Discussion

The results of the generation, annotation and thresholding procedures demonstrate the effectiveness of the proposed hybrid LLM-human approach in scaling the construction of a reliable adjectival hypernymy gold standard. Progressively refining the model-generated outputs, this process ensures that automatic inclusion criteria reflect the patterns of human acceptability.

After generation, the filtering and validation pipeline resulted in a final adjectival hypernymy dataset composed of 3,836 pairs, consisting of an average of 3.3 hypernyms per hyponym, considering that the original input hyponyms were 1,148 (see Section 4.1).

The large difference between categories regarding acceptance rate (see Table 1) reflects the progressive decrease in model agreement and reliability: consensus predictions are generally robust (87% accepted), majority ones require statistical validation but still lead to a significant inclusion rate (29%), while singletons mostly capture noisy or idiosyncratic associations (reflected by a negligible 0.6%).

The distinction in three sets and the analysis of model overlap show heterogeneous behaviour among models: gemma-3:27b tended to produce a

[11] The 2,433 pairs were reduced to 1,181 as in some cases there were multiple hyponyms for a candidate hypernym, so they were joined together in a merged list.

Error category	Example hypo-hyper pair	Singleton set (%)	Majority set (%)
No relation	*stupid-slow*	62	41
Similarity	*remarkable-extraordinary*	20	31
Directionality	*unknown-unfamiliar*	18	28

Table 3: Error analysis of models' outputs for the singleton and majority sets, reporting frequencies for the three error categories.

larger number of candidate hypernyms for each input hyponym (average of 9.44 across 4 folds), while mistral-small:24b displayed more conservative and repetitive generation patterns, with an overall Jaccard similarity of 0.76 across folds and an average of 3 candidates per hyponym. gpt-oss:20b offered an intermediate behaviour, balancing lexical productivity (4.59 candidates per hyponym) and confidence (similarity of 0.57), which is consistent with its lower proportion of singleton outputs (5%). The inter-model comparison confirms that ensemble prompting across multiple architectures increases coverage and provides opportunity to measure and scale the quality obtained. Candidates generated by at least two models were considerably more semantically coherent and better aligned with human judgments, confirming that cross-model consensus is a strong indicator of semantic validity.

The reverse hyponym generation applied to the singleton set functioned as an additional quality control measure to assess the semantic consistency of low-agreement outputs, resulting in a very low overall retention rate (7.4%). This elucidates once more how candidate hypernyms of the singleton set are not as reliable as others, confirmed by the fact that they were generated by only one of the models. On the one hand, this supports the hypothesis that hypernymy can be partially modeled as an entailment relation in the lexical space of adjectives, but on the other it also highlights the inherent difficulty of this task compared to noun-based hierarchies, more straightforward in both directions.

In order to more concretely assess this difference, a qualitative error analysis was performed along with the annotation of the two samples from the majority and singleton sets, individuating three main types of errors. A first category includes cases where models generated candidate hypernyms with meanings too deviating from the input hyponyms or not applicable to all contexts (e.g., *lively* as hypernym of *colourful*). A second class is related to semantic similarity: this involved adjectives that could be linked through a synonymy relation, being on the same semantic level, rather than hypernymy/hyponymy, being on two different hierarchical levels with one broader meaning including a more restricted one (e.g., *scattered* as hypernym of *distributed*). A third type is defined as opposite directionality: given an input hyponym adjective, the model would output a word with a narrower meaning, so a hyponym instead of a hypernym (e.g., *unfavorable* as hypernym of *bad*).

Among the annotated samples, those three main classes of errors were distributed as respectively 62%, 20% and 18% among the singleton refused pairs, and 41%, 31% and 28% among the majority ones (see Table 3). No cases of errors related to polysemy or PoS confusion were found, highlighting the role and importance of clear theoretical principles and definitions (see Section 3) in guiding models' predictions. As shown in Table 4, the inclusion of definitions in the final dataset and the integration with OEWN are essential, first to provide reliable semantic grounding to the methodology of this work, and second to introduce a valuable enrichment to this resource, aiming at modeling adjectival hypernymy links.

7. Conclusion

This study addresses the challenge of treatment and representation of hypernymy for adjectives, an aspect traditionally neglected in lexical resources such as WordNet. Through a hybrid approach combining different LLMs and human validation, it was possible to create an expanded and accurate dataset of hyponym-hypernym pairs for adjectives. Establishing rigorous theoretical guidelines ensured consistency in generation and annotation, improving reproducibility and reliability of results. Major contributions include creating a benchmark for evaluating hypernymy detection systems and enriching existing lexical resources, such as the OEWN.

Results show that the integration of emerging technologies with human control can accelerate the development of lexical resources while maintaining high quality standards. In addition, qualitative error analysis highlights the unique challenges posed by adjectives, offering insights for improving future semantic models.

8. Limitations and Future Work

Although the proposed methodology demonstrates scalability and general applicability, the present study has been conducted on a relatively small subset of highly-frequent English adjectives, not

Hypo	Hypo ID	Hypo definition	Hyper	Hyper ID	Hyper definition
evil	oewn-01134543-a	morally bad or wrong	*bad*	oewn-01129296-a	having undesirable or negative qualities
handsome	oewn-00220542-s	pleasing in appearance especially by reason of conformity to ideals of form and proportion	*beautiful*	oewn-00219320-a	delighting the senses or exciting intellectual or emotional admiration
incongruous	oewn-00564734-a	lacking in harmony or compatibility or appropriateness	*inappropriate*	oewn-00136789-a	not suitable for a particular occasion

Table 4: Examples of hyponym-hypernym pairs as appearing in the final dataset, containing the lemmas and their related OEWN synset IDs and definitions for each adjective.

containing rare or domain-specific adjectives that may display different relational behaviors. Moreover, despite the high IAA scores and acceptance rates achieved in manual and hybrid validation, a certain degree of subjectivity in human annotation remains inevitable, as judging hypernymy for adjectives often involves subtle interpretive decisions.

Hence, we open several directions for future research and resource development.

First, we aim to verify the portability of our hybrid LLM-human methodology to languages other than English, with two key objectives: on the computational side, this would allow to assess how LLMs perform in generating adjectival hypernyms for languages beyond English, including low-resource ones where lexical-semantic resources are sparse; on the linguistic side, it would offer the opportunity to examine whether and how the semantic organization of adjectives differs across languages, possibly providing new insights into cross-linguistic typologies and theoretical models of adjectival hierarchies.

Second, by extending the dataset, we plan to expand the OEWN with a large-scale integration of adjectival hyponym-hypernym pairs. While our released dataset already includes RDF links to OEWN synsets, our goal is to expand this coverage systematically across the full OEWN, creating a richer and more interconnected lexical resource that explicitly encodes adjectival hypernymy.

Finally, our evaluation set could be used in NLP applications and textual-lexical analysis, including improving semantic search, enhancing lexical entailment tasks, and refining contextual embeddings.

9. 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).

John P. McCrae is supported by Research Ireland under Grant Number SFI/12/RC/2289_P2 Insight_2, Insight SFI Centre for Data Analytics and Grant Number 13/RC/2106_P2, ADAPT SFI Research Centre.

10. Bibliographical References

Lorenzo Augello and John Philip McCrae. 2025. Inferring adjective hypernyms with language models to increase the connectivity of Open English Wordnet. In *Proceedings of the 5th Conference on Language, Data and Knowledge: The 5th OntoLex Workshop*, pages 1–11, Naples, Italy. Unior Press.

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10, Edinburgh, UK. Association for Computational Linguistics.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. SemEval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1081–1091, San Diego, California. Association for Computational Linguistics.

Zied Bouraoui, José Camacho-Collados, and Steven Schockaert. 2019. Inducing relational knowledge from BERT. *CoRR*, abs/1911.12753.

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa-Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. SemEval-2018 task 9: Hypernym discovery. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 712–724, New Or-

leans, Louisiana. Association for Computational Linguistics.

D. A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press, New York.

R. M. W. Dixon. 1982. *Where have All the Adjectives Gone?* De Gruyter Mouton, Berlin, New York.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Yoshihiko Hayashi. 2025. Evaluating LLMs' capability to identify lexical semantic equivalence: Probing with the word-in-context task. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6985–6998, Abu Dhabi, UAE. Association for Computational Linguistics.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, page 539–545, USA. Association for Computational Linguistics.

Frans Heyvaert. 2010. An outline for a semantic categorization of adjectives. In *Proceedings of the 14th EURALEX International Congress*, pages 1309–1318, Leeuwarden/Ljouwert, The Netherlands. Fryske Akademy.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Franz Hundsnurscher and Jochen Splett. 1982. *Grundlegung Einer Semantischen Beschreibung der Adjektive des Deutschen*, pages 16–47. VS Verlag für Sozialwissenschaften, Wiesbaden.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976, Denver, Colorado. Association for Computational Linguistics.

John Lyons. 1977. *Semantics*. Cambridge University Press.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plWordNet 3.0 – a comprehensive lexical-semantic resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2259–2268, Osaka, Japan. The COLING 2016 Organizing Committee.

John P. McCrae, Francesca Quattri, Christina Unger, and Philipp Cimiano. 2014. Modelling the semantics of adjectives in the ontology-lexicon interface. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, pages 198–209, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Viktor Moskvoretskii, Alexander Panchenko, and Irina Nikishina. 2024. Are large language models good at lexical semantics? a case of taxonomy learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1498–1510, Torino, Italia. ELRA and ICCL.

M. Lynne Murphy. 2003. *Semantic Relations and the Lexicon: Antonymy, Synonymy and other Paradigms*. Cambridge University Press.

Roberto Navigli, Paola Velardi, and Juana Maria Ruiz-Martínez. 2010. An annotated dataset for extracting definitions and hypernyms from the web. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open Dutch WordNet. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 302–310, Bucharest, Romania. Global Wordnet Association.

Victor Raskin and Sergei Nirenburg. 1995. Lexical semantics of adjectives a microtheory of adjectival meaning.

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102, Barcelona, Spain (Online). Association for Computational Linguistics.

Enrico Santus, Alessandro Lenci, Tin-Shing Chiu, Qin Lu, and Chu-Ren Huang. 2016. Nine features in a random forest to learn taxonomical semantic relations. In *Proceedings of the Tenth International Conference on Language Resources*

and Evaluation (LREC'16), pages 4557–4564, Portorož, Slovenia. European Language Resources Association (ELRA).

Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. 2015. EVALution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 64–69, Beijing, China. Association for Computational Linguistics.

Silke Scheible and Sabine Schulte im Walde. 2014. A database of paradigmatic semantic relation pairs for German nouns, verbs, and adjectives. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 111–119, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Martin Schmitt and Hinrich Schütze. 2021. Language models for lexical inference in context. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1267–1280, Online. Association for Computational Linguistics.

Yulia Tsvetkov, Nathan Schneider, Dirk Hovy, Archna Bhatia, Manaal Faruqui, and Chris Dyer. 2014. Augmenting English adjective senses with supersenses. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4359–4365, Reykjavik, Iceland. European Language Resources Association (ELRA).

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, Online. Association for Computational Linguistics.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. HyperLex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4):781–835.

Ivan Vulić and Nikola Mrkšić. 2018. Specialising word vectors for lexical entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1134–1145, New Orleans, Louisiana. Association for Computational Linguistics.

11. Language Resource References

BNC Consortium. 2007. *British National Corpus 1994*. Literary and Linguistic Data Service.

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Kilgarriff, Adam. 1997. *Putting frequencies in the dictionary*.

McCrae, John P. and Rademaker, Alexandre and Bond, Francis and Rudnicka, Ewa and Fellbaum, Christiane. 2019. *English WordNet 2019 – An Open-Source WordNet for English*. Global Wordnet Association.

Miller, George A. 1995. *WordNet: a lexical database for English*. Association for Computing Machinery.