# Large Language Models for Lexical Resource Enhancement: Multiple Hypernymy Resolution in WordNet

Dimitar Hristov
Institute for Bulgarian Language
Bulgarian Academy of Sciences
dimitar@dcl.bas.bg

## Abstract

Large language models (LLMs) have materially changed natural language processing (NLP). While LLMs have shifted focus from traditional semantic-based resources, structured linguistic databases such as WordNet remain essential for precise knowledge retrieval, decision making and aiding LLM development. WordNet organizes concepts through synonym sets (synsets) and semantic links but suffers from inconsistencies, including redundant or erroneous relations. This paper investigates an approach using LLMs to aid the refinement of structured language resources, specifically WordNet, by an automation for multiple hypernymy resolution, leveraging the LLMs semantic knowledge to produce tools for aiding and evaluating manual resource improvement.

## 1 Introduction

In recent years, an acceleration in the development of AI, machine learning and specifically generative models have greatly expanded the capabilities for solving tasks in the field of natural language processing (NLP). Large language models have proven to be a powerful tool for word sense disambiguation, sentiment analysis, abstractive summarization, paraphrasing with sentiment change, and other tasks.

The focus in natural language processing has in large part been shifted from development of structured language resources to the now more popular large language models. LLMs, which are themselves not just language resources but powerful often general-purpose tools, allow easy adaptability and specialization through fine tuning and prompt engineering.

There are, however, two reasons for the continued development of structured data resources. Such structured language data resources include the various forms of dictionaries - entry-based data with predefined parts such as word, inflection, definitions, examples - as well as ontology-based resources like WordNet (Miller et al., 1990; Fellbaum, 1998), BalkaNet (Tufis et al., 2004) and EuroWordNet (Vossen, 1998), incorporating vast amounts of knowledge with high accuracy. First and key in many spheres such as medicine and biology, structured data resources are deterministic, precise and validated. This ensures decisions are made on a consistent and provably correct data. This contrasts to the results from LLMs, where hallucinations - factually or logically unsound responses - occur to often for the extracted information to be readily usable without additional validation in high-stakes environments.

Additionally, this same power of LLMs is based on a very large preexisting knowledge base, which is incorporated in the model through training and fine-tuning. Existing structured language data resources are a significant knowledge-baring part of the training corpus for LLMs, meaning their continued development is essential for the progression of large language models. Even then, not all ontologies and knowledge bases have been used for model training.

The aim of this paper is to explore the viability of LLMs as tool for aiding and evaluation of structured language data enhancement with a focus on WordNet.

### 1.1 WordNet and multiple hypernymy

WordNet is an ontology-based structured language resource aiming to represent the interconnectedness of language concepts by constructing a network of concepts represented

*Proceedings of the Student Research Workshop associated with RANLP-2025*, pages 20–26, Varna, Bulgaria, Sep 8-10, 2025.
https://doi.org/10.26615/issn.2603-2821.2025_003

by synonym sets or synsets – sets of words or multi-word expressions with a common meaning – and the various semantic relationships between them. The resource has a graph-based structure well suited for deterministic approaches in NLP task solving.

One of the key ideas of WordNet is the codification of inheritance as hypernymy and hyponyny relations, linking a more general to a more specific concept - concept A (hyponym) is a type of concept B (hypernym), e.g., {bee:1} is a type of {hymenopterous insect:1; hymenopteran:1; hymenopteron:1; hymenopter:1}, which is itself a type of {insect:1}.

As any manually created database of knowledge, differences of language perception, ambiguity and other factors may occasionally cause errors in both the lexical data and the structure within WordNet (Richens, 2008; Verdezoto and Vieu, 2011). These can include: missing or erroneous words in the synset, errors in definition, synset ambiguity (one synset representing multiple concepts), multiple synsets for the same concept, wrong relation types, missing relations. Koeva and Hristov (2023) define one such potential issue - erroneous or extra hypernyms where no or other relations should be. They give a manually crafted dataset with resolved multiple hypernymy, resulting in a tree hypernymy structure, which requires further evaluation.

This paper will test whether the process of resolving multiple hypernymy can be automated through the use of LLMs and prompt engineering, evaluate the results and propose uses for LLMs in the WordNet improvement process.

### 1.2 Paper outline

Section 1 introduced the context and aim of the paper, while Section 2 links to the base research on which the task is defined. The methodology of the experiment, data and implementation are described in Section 3. Section 4 analyses the outputs and measurements of the results with a proposal for uses of the setup. Section 5 explores a list of potential improvements and extentions of the current work.

## 2 Related work

This paper looks into an approach to automate an otherwise manual task related to the creation and maintenance of structured language resources. In the particular task chosen for the experiment, the automated task is connected to the nature of hypernymy relations between synsets and their validity. A manual execution of multiple hypernymy resolution has been performed by Koeva and Hristov (2023) with promising results, invoking a question on whether such phenomena can be evaluated and modified in an automated or semi-automated way.

Lippolis et al. (2025) explore the automatic construction of an ontology draft using subtask-decomposed prompting, as well as prompting technique based on Chain Of Thought (CoT), where LLM inference is done separately on atomic data point - in this case competence questions, later merge together in a full ontology. A similar approach of dividing the problem into per-unit tasks is taken within the current work.

## 3 Methodology

The aim of the paper is to evaluate the effectiveness and efficacy of LLMs as a tool to aid with WordNet structural enhancement. This was achieved through emulating a standard workflow - solving hypernymy resolution tasks separately in a series.

### 3.1 Structure

The experiment is structured as a series of instruction-based multiple-choice tasks. The experiment is performed with generic out-of-the-box LLMs without any additional task-specific training or fine tuning. An inference is run for each separate synset with multiple hypernymy, using a prompt as described in A which provides:

1. General instructions - LLM's role (WordNet expert), task context (synsets and hypernymy relations), input format (how synset data is provided) and output format (a single synset ID);

2. Examples - this part is optional and is either missing (A.1 0-shot), or provides 1 or 5 examples (A.2 1-shot or few-shot);

3. The main task - a list of the current hypernym synsets, the question (Which synset above is the best hypernym?) and a description of the synset for which a hypernym is to be chosen.

### 3.2 Data

The experiment uses data on synset words, relations and meaning from Princeton WordNet 3.0. The data set was filtered to include only details on synsets with two or more hypernyms - a total of 1421 synsets - their word lists, meanings and hypernymy relations. As evaluation was done using the resulting data from Koeva and Hristov (2023), the data was synchronized, leaving only those synsets for which one of the already existing hypernyms was selected. Koeva and Hristov (2023) assigned a new hypernym to 77 of the synsets. Additionally, 5 synsets were selected to be used as examples in 1-shot and few-shot prompts, leaving 1339 synsets for evaluation.

The five manually chosen examples are:

1. Hyponym {mathematical space:1; topological space:1} "(mathematics) any set of points that satisfy a set of postulates of some kind" with hypernyms:
   - {space:1; infinite:2} "the unlimited expanse in which everything is located"
   - {set:41} "(mathematics) an abstract collection of numbers or symbols"
   - Chosen hypernym: {set:41}
2. Hyponym {Calamagrostis:1; genus Calamagrostis:1} "reed grass" with hypernyms:
   - {monocot genus:1; liliopsid genus:1} "genus of flowering plants having a single cotyledon (embryonic leaf) in the seed"
   - {genus:2} "(biology) taxonomic group containing one or more species"
   - Chosen hypernym: {monocot genus:1; liliopsid genus:1}
3. Hyponym {altar boy:1} "a boy serving as an acolyte" with hypernyms:
   - {acolyte:1} "someone who assists a priest or minister in a liturgical service; a cleric ordained in the highest of the minor orders in the Roman Catholic Church but not in the Anglican Church or the Eastern Orthodox Churches"
   - {male child:1; boy:3} "a youthful male person"
   - Chosen hypernym: {male child:1; boy:3}
4. Hyponym {potato:1; white potato:1; Irish potato:1; murphy:1; spud:4; tater:1} "an edible tuber native to South America; a staple food of Ireland" with hypernyms:
   - {starches:1} "foodstuff rich in natural starch (especially potatoes, rice, bread)"
   - {solanaceous vegetable:1} "any of several fruits of plants of the family Solanaceae; especially of the genera Solanum, Capsicum, and Lycopersicon"
   - {root vegetable:1} "any of various fleshy edible underground roots or tubers)"
   - Chosen hypernym: {solanaceous vegetable:1}
5. Hyponym {water:6} "a liquid necessary for the life of most animals and plants" with hypernyms:
   - {food:1; nutrient:1} "any substance that can be metabolized by an animal to give energy and build tissue"
   - {nutrient:2} "any substance (such as a chemical element or inorganic compound) that can be taken in by a green plant and used in organic synthesis"
   - {liquid:11} "a substance that is liquid at room temperature and pressure"
   - Chosen hypernym: {liquid:11}

### 3.3 Implementation

The experiment was implemented using scripts written in bash script or Python, Ollama[1] for local inference execution and the LangChain framework[2] with the LangChain Ollama

[1] https://ollama.com/

[2] https://www.langchain.com/

integration library for the application. Inference was done on four widely available LLMs - Google Gemma 3 with 4 billion parameters, Meta Llama 3.1 with 8 billion parameters, Mistral with 7 billion parameters and Microsoft Phi-4 with 14 billion parameters.

The models were retrieved from the Ollama model library as 4-bit quantized. The temperature (creativeness) setting was set to 0.7, while the number of examples was varied between none for zero-shot execution, 1 for 1-shot execution and 5 for few-shot execution, resulting in a total of 12 runs. In cases where the inference execution returned an invalid response, i.e., not a well-formatted synset ID or not the ID of one of the given hypernym synsets, up to two additional inferences were performed for the specific synset.

The code, data and generated results are available on GitHub[3].

## 4 Results and evaluation

The main measure used for the evaluation of the results from running the experiment was agreement - the ratio of synsets, for which an LLM has assigned the same hypernym as set in the manual dataset, or the ratio of synsets for which two LLMs have assigned the same hypernym. This measure shows generally whether LLMs' probabilistic generation can emulate a human's logic, and whether a confidence measure can be established for the LLM's results. All measurements are presented in Appendix B Agreement tables.

Tables 1, 2 and 3 present the agreement measure between each individual LLM and the manual dataset, as well as between each 2 LLMs. The measurements for agreement with the manual hypernymy resolution range between 45% and 55% regardless of number of examples or LLM, suggesting that no correlation is present between the manual approach and the LLM inference. However, the agreement between LLMs is consistently higher at 52.7-71.5% for 0-shot, 63.9-77.6% for 1-shot and 62.0-75.8% for few-shot. This suggests that (1) examples improves the understanding of the task, leading to more consistent results from LLMs, and (2) different LLMs may have more similar training data, most certainly all containing WordNet knowledge in addition to other publicly available datasets, while a human possesses different and additional knowledge, causing the consistency between LLMs and no apparent correlation between LLM results and the manual resolution.

Table 4 presents the ratio of synset hypernyms assignments for which there is a majority opinion - at least 3 of the 4 LLMs have proposed the same assignment. The results show Gemma 3 as an outlier, with participation in the majority for 66.8% of synsets for 1-shot inference, while other models agree with the majority for 76.9-79.1% of synsets. Tables 5 and 6 present the combined agreement of the LLMs with the manual data where (1) at least 3 LLMs have produced the same output, and (2) where all 4 LLMs have proposed the same resolution (unanimity). These tables show a potential for:

- using LLMs as a starting point for aided manual performance of hypernymy resolution, with a promising 47.9% unanimity, 36.3% non-unanimous majority and only 15.7% without agreement;

- using LLMs as an evaluation tool for a performed manual hypernymy resolution, focusing attention on cases where LLMs have a unanimous (22.8% for 1-shot) ana non-unanimous majority (19.7%) disagreement with the manual results.

## 5 Further work

The evaluation of the results of this study provide an overview of the potential use of LLMs in the improvement of structured language data resources. Several improvements can be made in the experiment to ensure consistency and validity of the model responses:

- addition of more and diverse LLMs - this will give more weight and granularity to the agreement measure;

- grouping of synsets by category, yielding more consistent logic with added information for the task;

- addition of human evaluation for both the original proposed resolution by Koeva and Hristov (2023) and the LLM results;

[3] https://github.com/DCL-IBL/SemNet

Kaplan and Schubert (2001); Gangemi et al. (2001) note that multiple hypernymy often encode other relation types, a case for further WordNet structure modifications. Koeva and Hristov (2023) explore this extension to the multiple hypernymy resolution - resolution of alternative relation types for existing hypernyms. This may be an additional target for LLM-aided enhancement and evaluation, using an improved variant of the experiment setup.

## Acknowledgments

The present study is carried out within the project Infrastructure for Fine-tuning Pre-trained Large Language Models, Grant Agreement No. ПВУ - 55 from 12.12.2024 /BG-RRP-2.017-0030-C01/.

## References

Christiane Fellbaum, editor. 1998. WordNet: An Electronic Lexical Database. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Aldo Gangemi, Nicola Guarino, and Alessandro Oltramari. 2001. Conceptual analysis of lexical taxonomies: The case of wordnet top-level. CoRR, cs.CL/0109013.

Aaron Kaplan and Lenhart Schubert. 2001. Measuring and improving the quality of world knowledge extracted from wordnet.

Svetla Koeva and Dimitar Hristov. 2023. Resolving multiple hyperonymy. In Proceedings of the 12th Global Wordnet Conference, pages 343–351, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Anna Sofia Lippolis, Mohammad Javad Saeedizade, Robin Keskisärkkä, Sara Zuppiroli, Miguel Ceriani, Aldo Gangemi, Eva Blomqvist, and Andrea Giovanni Nuzzolese. 2025. Ontology generation using large language models.

George A. Miller, Richard Beckwith, Christiane. Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to WordNet: An On-line Lexical Database. International journal of lexicography, 3(4):235–244.

Tom Richens. 2008. Anomalies in the wordnet verb hierarchy. pages 729–736.

Dan Tufis, Dan Cristea, and Sofia Stamou. 2004. Balkanet: Aims, methods, results and perspectives – a general overview. Romanian Journal of Information Science and Technology Special Issue, 7:9–43.

Nervo Verdezoto and Laure Vieu. 2011. Towards semi-automatic methods for improving wordnet. In Proceedings of the Ninth International Conference on Computational Semantics, IWCS '11, page 275–284, USA. Association for Computational Linguistics.

Piek Vossen, editor. 1998. EuroWordNet: a multilingual database with lexical semantic networks. Kluwer Academic Publishers, USA.

# A Prompts

## A.1 0-shot

You are a WordNet expert. Your task is to evaluate hypernymy relations between semantic concepts. Each semantic concept is represented by a group of words with common meaning. This group is called a synset. If concept A is a hypernym of concept B, then concept B is a type of concept A, and concept A is a more generic version of concept B.

Each synset is presented by its ID, group of words and meaning. You will be given a synset and its hypernyms and will be instructed to choose a single hypernym.

Reply only with the chosen hypernym synset ID with format 30-<8 digits>-n and no other words. Do not give any reasoning and do not generate other text.

You are given the following synsets:
- ID $(ID_a)$ with words $(words_a)$ and meaning $(definition_a)$
...
- ID $(ID_x)$ with words $(words_x)$ and meaning $(definition_x)$

Which of the synsets $(ID_a)$... and $(ID_x)$ is most likely to be the hypernym of synset $(ID_{hypo})$ defined as:
- ID $(ID_{hypo})$ with words $(words_{hypo})$ and meaning $(definition_{hypo})$

## A.2 1-shot or few-shot

You are a WordNet expert. Your task is to evaluate hypernymy relations between semantic concepts. Each semantic concept is represented by a group of words with common meaning. This group is called a synset. If concept A is a hypernym of concept B, then concept B is a type of concept A, and concept A is a more generic version of concept B.

Each synset is presented by its ID, group of words and meaning. You will be given a synset and its hypernyms and will be instructed to choose a single hypernym.

Reply only with the chosen hypernym synset ID with format 30-<8 digits>-n and no other words. Do not give any reasoning and do not generate other text.

EXAMPLE [$(n)$]

You are given the following synsets:
- ID $(ID_a^{ex.n})$ with words $(words_a^{ex.n})$ and meaning $(definition_a^{ex.n})$
...
- ID $(ID_x^{ex.n})$ with words $(words_x^{ex.n})$ and meaning $(definition_x^{ex.n})$

Which of the synsets $(ID_a^{ex.n})$... and $(ID_x^{ex.n})$ is most likely to be the hypernym of synset $(ID_{hypo}^{ex.n})$ defined as:
- ID $(ID_{hypo}^{ex.n})$ with words $(words_{hypo}^{ex.n})$ and meaning $(definition_{hypo}^{ex.n})$

$(ID_{result}^{ex.n})$

...

TASK

You are given the following synsets:
- ID $(ID_a)$ with words $(words_a)$ and meaning $(definition_a)$
...
- ID $(ID_x)$ with words $(words_x)$ and meaning $(definition_x)$

Which of the synsets $(ID_a)$... and $(ID_x)$ is most likely to be the hypernym of synset $(ID_{hypo})$ defined as:
- ID $(ID_{hypo})$ with words $(words_{hypo})$ and meaning $(definition_{hypo})$

## B Agreement tables

| 0-shot | Manual | Gemma 3 4B | Llama 3.1 8B | Mistral 7B | Phi-4 14B |
|---|---|---|---|---|---|
| Manual | - | 45.4% | 50.4% | 50.9% | 54.1% |
| Gemma 3 4B | 45.4% | - | 55.4% | 52.7% | 57.6% |
| Llama 3.1 8B | 50.4% | 55.4% | - | 71.5% | 70.9% |
| Mistral 7B | 50.9% | 52.7% | 71.5% | - | 64.3% |
| Phi-4 14B | 54.1% | 57.6% | 70.9% | 64.3% | - |

Table 1: Measures of agreement between LLMs and manual resolution for runs without examples

| 1-shot | Manual | Gemma 3 4B | Llama 3.1 8B | Mistral 7B | Phi-4 14B |
|---|---|---|---|---|---|
| Manual | - | 53.3% | 49.2% | 48.7% | 48.9% |
| Gemma 3 4B | 53.3% | - | 67.1% | 65.8% | 63.9% |
| Llama 3.1 8B | 49.2% | 67.1% | - | 77.6% | 76.9% |
| Mistral 7B | 48.7% | 65.8% | 77.6% | - | 76.4% |
| Phi-4 14B | 48.9% | 63.9% | 76.9% | 76.4% | - |

Table 2: Measures of agreement between LLMs and manual resolution for runs with 1 example

| Few-shot | Manual | Gemma 3 4B | Llama 3.1 8B | Mistral 7B | Phi-4 14B |
|---|---|---|---|---|---|
| Manual | - | 51.0% | 47.4% | 50.1% | 49.2% |
| Gemma 3 4B | 51.0% | - | 62.0% | 63.6% | 58.6% |
| Llama 3.1 8B | 47.4% | 62.0% | - | 75.8% | 71.9% |
| Mistral 7B | 50.1% | 63.6% | 75.8% | - | 69.7% |
| Phi-4 14B | 49.2% | 58.6% | 71.9% | 69.7% | - |

Table 3: Measures of agreement between LLMs and manual resolution for runs with 5 examples

| Majority | At least 3 LLMs agree | Gemma 3 4B | Llama 3.1 8B | Mistral 7B | Phi-4 14B |
|---|---|---|---|---|---|
| 0-shot | 69.9% | 56.9% | 64.4% | 60.6% | 62.1% |
| 1-shot | 84.3% | 66.8% | 79.1% | 78.1% | 76.9% |
| 5-shot | 81.6% | 63.0% | 75.7% | 75.4% | 70.9% |

Table 4: Measures of existence and LLM agreement with majority

| Manual | Majority (at least 3 LLMs) | Manual agrees | Manual disagrees | No majority |
|---|---|---|---|---|
| 0-shot | 69.9% | 37.7% | 32.2% | 30.1% |
| 1-shot | 84.3% | 41.8% | 42.5% | 15.7% |
| 5-shot | 81.6% | 40.8% | 40.8% | 18.4% |

Table 5: Measures of agreement of manual results with majority

| Manual | Unanimity (all 4 LLMs) | Manual agrees | Manual disagrees | No unanimity |
|---|---|---|---|---|
| 0-shot | 34.3% | 18.9% | 15.4% | 65.7% |
| 1-shot | 47.9% | 25.2% | 22.8% | 52.1% |
| 5-shot | 40.3% | 19.7% | 20.5% | 59.7% |

Table 6: Measures of agreement of manual results with majority