# Measuring and Comparing Word Similarity Using WordNet

## A Project in Computational Linguistics

K S Saisankalp Davey
Yash More
UG-2 CLD

**IIIT - Hyderabad**



Computational Linguistics - 2

December 5, 2025

# Contents

**Abstract**

This report summarizes the evaluation of four classical WordNet-based semantic similarity measures—Extended Lesk, Leacock-Chodorow (LCH), Shortest-Path, and Wu-Palmer (WUP)—on two complementary benchmarks (SimLex-999 and WordSim-353). We provide tabular results, visual comparisons, and qualitative observations derived from the relationship-type analysis conducted on the shared datasets.

# 1 Introduction

Click here to visit the github repo for the project. This project examines how classical WordNet-based similarity measures approximate human judgements along three research questions:

1. benchmark the shortest-path, Leacock–Chodorow (LCH), Wu–Palmer (WUP), and Extended Lesk algorithms against standard similarity/relatedness datasets
2. diagnose which lexical relationship types—synonyms, hypernyms, functional associations, topic links, and others—each measure captures reliably versus where they fail

The present report delivers the comparative and relationship-type analyses using the available results while noting that the full context-aware evaluation remains outstanding because the original SCWS benchmark could not be found as the web pages hosting the benchmark were not accessible.

# 2 Experimental Setup

## 2.1 Datasets

- **SimLex-999**: 999 carefully curated word pairs emphasizing pure similarity. POS tags (nouns, verbs, adjectives) enable POS-specific diagnostics.
- **WordSim-353**: 353 word pairs reflecting broader semantic relatedness. No explicit POS labels were provided, so all measures treated pairs as candidate nouns/verbs.

## 2.2 Algorithms

- **Extended Lesk**: Overlap across extended glosses (definitions, examples, related synsets) with geometric normalization.
- **Leacock-Chodorow (LCH)**: Path-based similarity using maximum taxonomy depth; restricted to noun/verb hierarchies.
- **Shortest-Path**: Raw inverse path-length between synsets, normalized by taxonomy depth.
- **Wu-Palmer (WUP)**: Similarity estimated from the depth of the least common subsumer relative to synset depths.

All evaluations relied on the WordNet 3.0 taxonomy via NLTK. Coverage differences arise from inherent algorithmic constraints.

# 3 Results

## 3.1 Overall Correlations

Following tables present the headline correlation and coverage metrics:

Table 1: SimLex-999 overall performance.

| Algorithm | Spearman $\rho$ | Pearson $r$ | Coverage (%) |
|---|---|---|---|
| Extended Lesk | 0.4511 | 0.4399 | 98.3 |
| LCH | 0.5822 | 0.5821 | 81.1 |
| Shortest-Path | 0.4561 | 0.4736 | 100.0 |
| Wu-Palmer | 0.4389 | 0.4327 | 100.0 |

Table 2: WordSim-353 overall performance.

| Algorithm | Spearman $\rho$ | Pearson $r$ | Coverage (%) |
|---|---|---|---|
| Extended Lesk | 0.5021 | 0.5532 | 94.9 |
| LCH | 0.4124 | 0.5125 | 99.4 |
| Shortest-Path | 0.3965 | 0.5484 | 100.0 |
| Wu-Palmer | 0.4423 | 0.4567 | 100.0 |

## 3.2 Visual Comparison

Following plots summarize rank correlation, linear correlation, and coverage respectively.

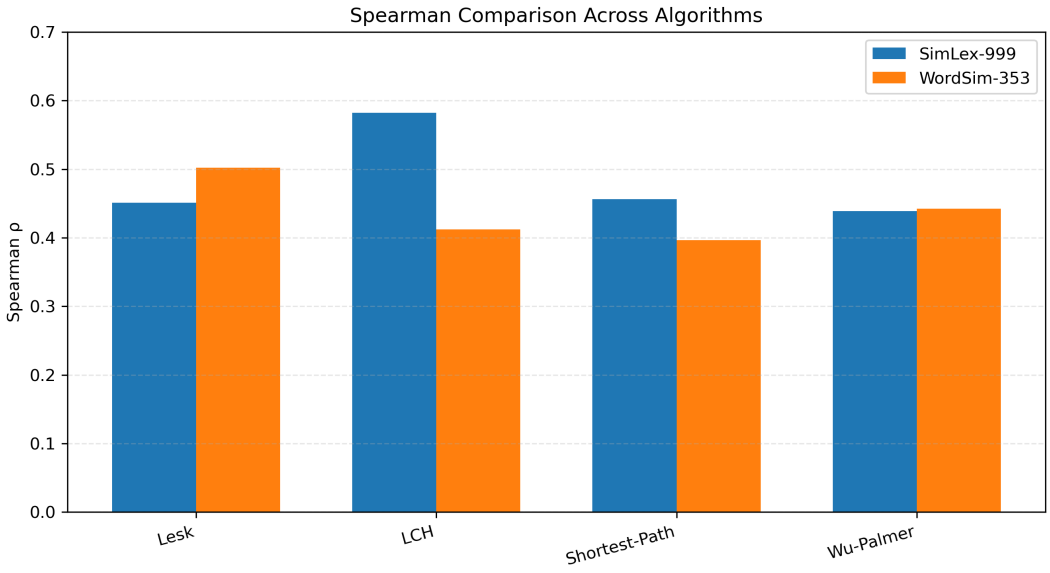


Figure 2: Spearman correlation comparison across algorithms and datasets.

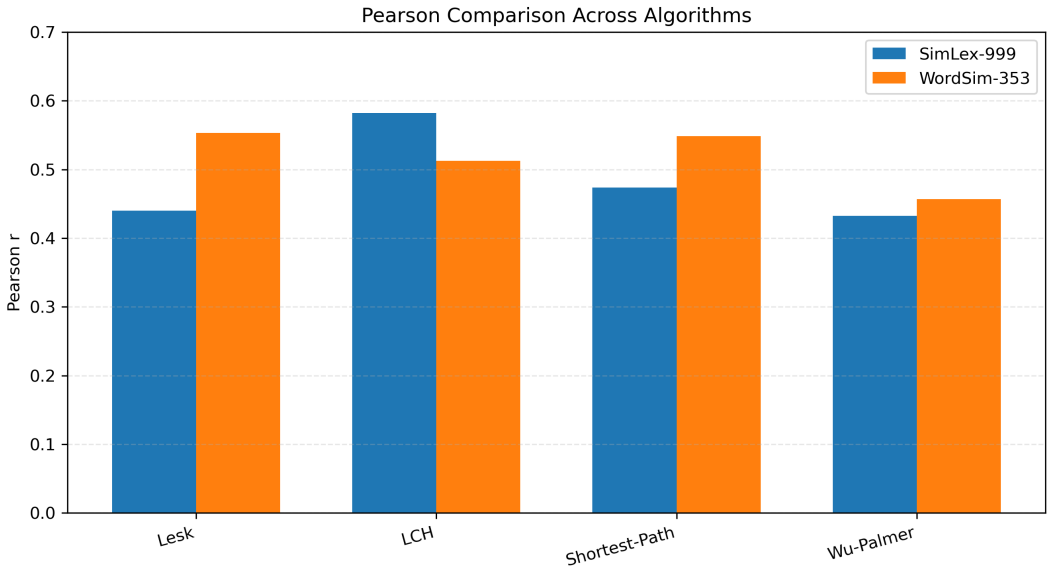


Figure 3: Pearson correlation comparison across algorithms and datasets.

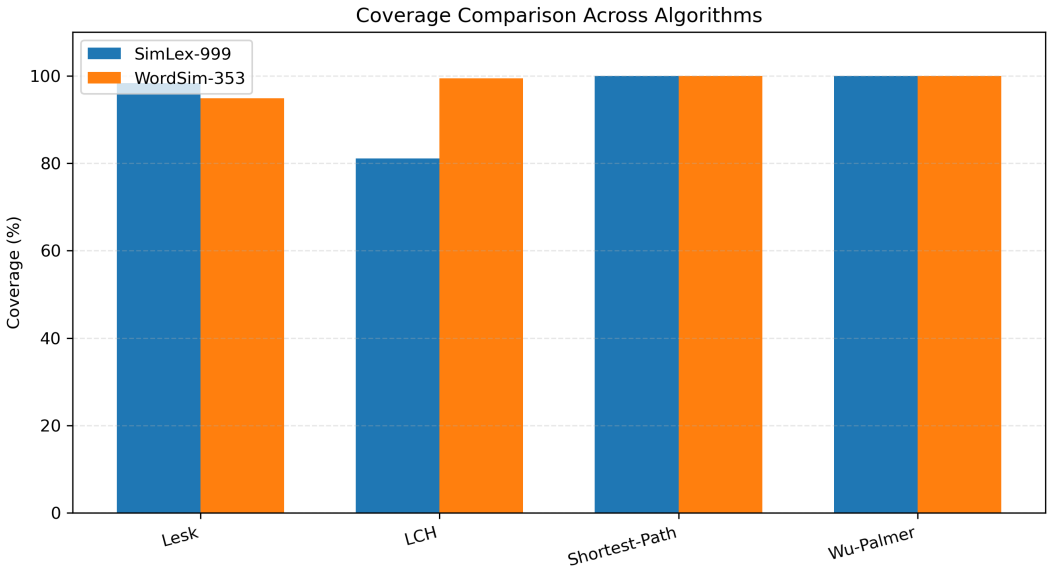


Figure 4: Coverage comparison (percentage of evaluated pairs).

### 3.3 Performance by Relationship Type

To make the behaviour of the four measures more comparable across semantic relation types, Tables 3 and 4 summarize Spearman correlations and coverage for each algorithm within the manually assigned relationship categories on SimLex-999 and WordSim-353 respectively.

Table 3: Performance by relationship type on SimLex-999. Spearman $\rho$ and coverage (%). "const." indicates near-constant predictions (undefined correlation).

| Relationship | Extended Lesk | LCH | Shortest-Path | Wu–Palmer |
|---|---|---|---|---|
| Antonym | $\rho = 0.32$, 100% | $\rho = 0.14$, 50% | $\rho = 0.48$, 100% | $\rho = 0.37$, 100% |
| Synonym | const., 100% | $\rho = -0.11$, 79% | const., 100% | $\rho = -0.14$, 100% |
| Hypernym–Hyponym | $\rho = 0.17$, 100% | $\rho = 0.18$, 100% | $\rho = 0.21$, 100% | $\rho = 0.06$, 100% |
| Co-hyponym | $\rho = -0.12$, 100% | $\rho = 0.17$, 100% | const., 100% | $\rho = 0.15$, 100% |
| Functional | $\rho = 0.12$, 98% | $\rho = 0.29$, 84% | $\rho = 0.21$, 100% | $\rho = 0.13$, 100% |
| Meronym–Holonym | $\rho = 0.30$, 100% | $\rho = -0.26$, 100% | $\rho = -0.26$, 100% | $\rho = -0.57$, 100% |
| Unrelated | $\rho = 0.52$, 95% | $\rho = 0.18$, 42% | $\rho = -0.31$, 100% | $\rho = 0.37$, 100% |

Table 4: Performance by relationship type on WordSim-353. Spearman $\rho$ and coverage (%). "const." indicates near-constant predictions (undefined correlation); "N/A" for categories with too few pairs.

| Relationship | Extended Lesk | LCH | Shortest-Path | Wu–Palmer |
|---|---|---|---|---|
| Antonym | N/A | N/A | N/A | N/A |
| Synonym | const., 100% | const., 100% | const., 100% | const., 100% |
| Hypernym–Hyponym | $\rho = 0.63$, 98% | $\rho = 0.62$, 100% | $\rho = 0.62$, 100% | $\rho = 0.68$, 100% |
| Co-hyponym | $\rho = -0.16$, 100% | const., 100% | const., 100% | $\rho = 0.20$, 100% |
| Functional | $\rho = 0.28$, 96% | $\rho = 0.18$, 100% | $\rho = 0.18$, 100% | $\rho = 0.24$, 100% |
| Meronym–Holonym | $\rho = 1.00$, 100% | $\rho = -0.50$, 100% | $\rho = -0.50$, 100% | $\rho = -0.50$, 100% |
| Unrelated | $\rho = 0.24$, 80% | $\rho = -0.11$, 95% | $\rho = -0.16$, 100% | $\rho = -0.10$, 100% |

### 3.4 Relationship-Type Analysis

Beyond aggregate correlations, we manually categorized SimLex-999 and WordSim-353 pairs into coarse semantic relationship types (synonyms, hypernym-hyponym, co-hyponyms, antonyms, meronyms-holonyms, functional-associative links, and loosely topic-related pairs) using WordNet structure and gloss cues. This allows us to see *what kind* of lexical knowledge each algorithm captures, rather than only how well it fits the global ranking.

**Extended Lesk:** Lesk, which relies on gloss and example overlap, behaves most stably on pairs that are either clearly unrelated or broadly topic-linked. For SimLex, it achieves its highest rank correlation on the "unrelated" category (approximately $\rho \approx 0.52$), correctly assigning low scores to most genuinely dissimilar pairs. On WordSim, Lesk also tracks many functional or topical associations (e.g., *doctor–hospital*, *coffee–cup*) reasonably well, since these concepts frequently co-occur in WordNet definitions and examples. However, for tight synonym and co-hyponym pairs, the extended glosses of different senses often share many tokens, leading to near-constant high scores across multiple pairs and hence reduced rank discrimination. This "gloss-overlap saturation" explains why synonym blocks can be scored as uniformly similar, even when human ratings show subtle gradations.

**Leacock–Chodorow (LCH):** LCH, being purely path-based, is most effective on pairs that sit in a clean taxonomic chain. In both datasets it achieves its best behaviour on hypernym–hyponym

relations (e.g., *dog–animal*, *car–vehicle*), with moderate to strong positive correlations ($\rho \approx 0.18$ on SimLex hypernyms, $\rho \approx 0.62$ on WordSim hypernyms). Co-hyponyms (sisters under the same parent) are also handled better than chance, although their similarity is often underestimated because the measure only sees path length, not shared functional roles. LCH performs poorly on meronymy and holonymy (part–whole) relations and on many verb antonyms: WordNet's noun/verb hierarchies do not encode part-of or oppositeness via short IS–A paths, so these pairs either receive low similarity or are dropped due to missing paths. Adjectives and adverbs are effectively unsupported, which directly limits coverage on SimLex adjective pairs.

**Shortest-Path:** The shortest-path measure makes strong use of the global taxonomy and therefore attains near-full coverage on both benchmarks. Its behaviour across relationship types closely mirrors that of LCH, but with smoother score distributions: hypernym–hyponym pairs are again the easiest and exhibit positive correlations, while meronym/holonym and loosely associated pairs often receive noisy or even negatively correlated scores. The measure also tends to underestimate the highest-similarity synonym and near-synonym pairs—especially when the two words belong to slightly different but closely aligned subtrees (e.g., *shore–coast*, *intelligent–smart*)—because similarity is strictly bounded by path length in the taxonomy rather than by shared usage or topical proximity.

**Wu–Palmer (WUP):** The Wu–Palmer measure similarly exploits the full taxonomy and achieves near-complete benchmark coverage, but it incorporates least-common-subsumer depth in addition to path length, yielding more graded similarity scores. As with shortest-path and LCH, hypernym–hyponym relations receive the most consistent positive correlations, whereas meronym/holonym and weak associative pairs remain unstable and occasionally negatively correlated. Despite its depth normalisation, WUP still systematically underestimates strong synonymy and near-synonymy across adjacent subtrees (e.g., *shore–coast*, *intelligent–smart*), since its upper bound is imposed by taxonomic structure rather than distributional or usage-based similarity.

**Summary:** Overall, the relationship-type breakdown confirms a complementary pattern: structural, path-based methods (Shortest-Path, LCH, WUP) are well matched to hierarchical IS–A relations but blind to many associative or part-whole links, while gloss-based Lesk is more forgiving for functionally or topically related concepts but struggles to finely rank tightly clustered synonym and co-hyponym sets. These strengths and weaknesses are consistent with the underlying information each method exploits (taxonomy vs. glosses) and help explain why no single measure dominates across all pair types.

### 3.5 Relationship-Type Visualizations

To visualize the differential performance of each algorithm across relationship categories, we present heatmaps (Figures 5 and 6), grouped bar charts (Figures 7 and 8), and a faceted comparison (Figure 9) that juxtaposes both datasets side-by-side for common relationship types.

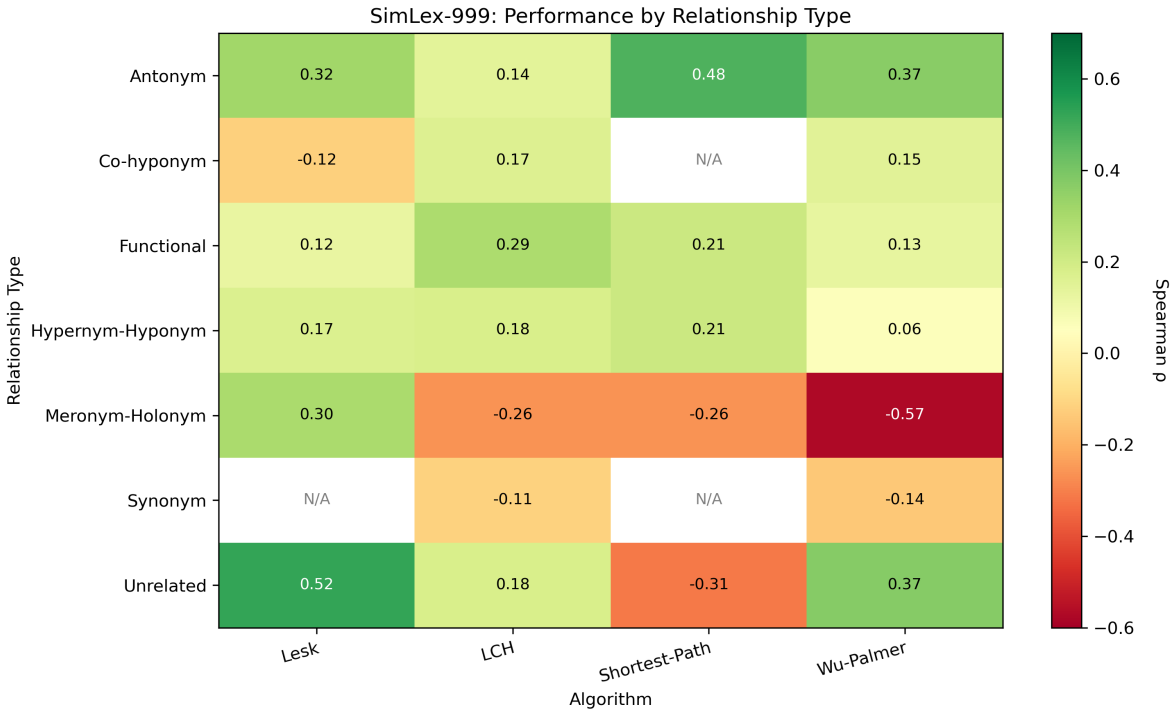


Figure 5: Heatmap of Spearman $\rho$ by relationship type on SimLex-999. Green indicates positive correlation, red negative, and gray denotes undefined (constant predictions).

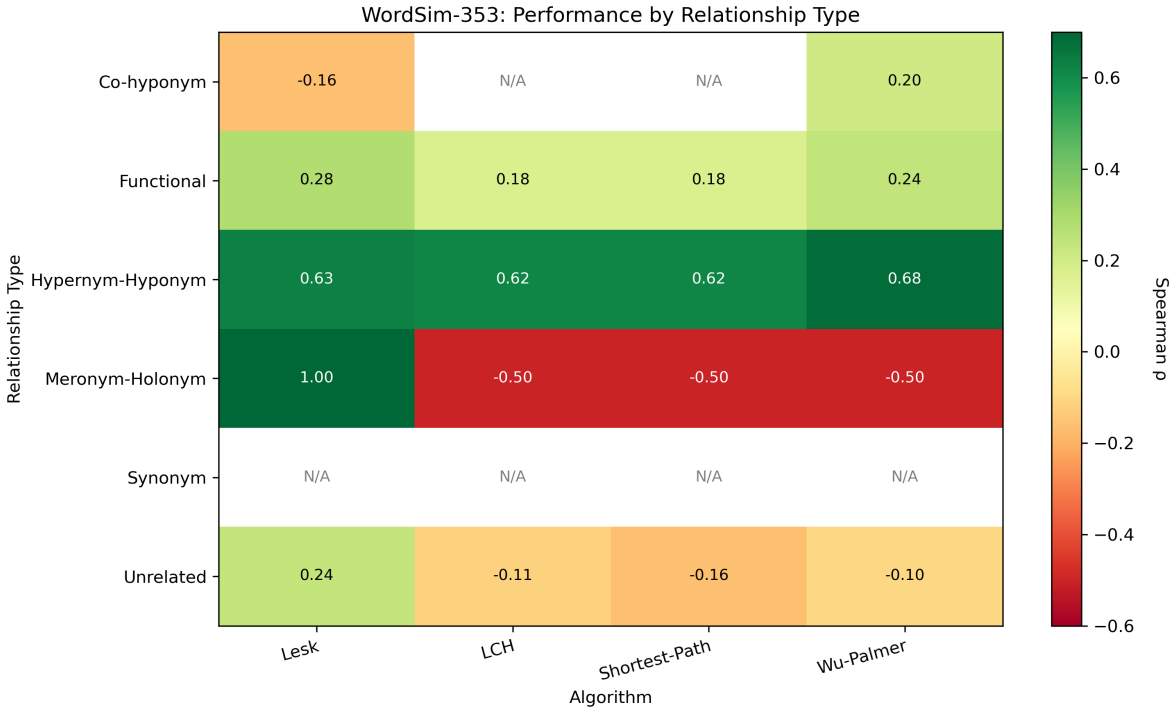


Figure 6: Heatmap of Spearman $\rho$ by relationship type on WordSim-353.

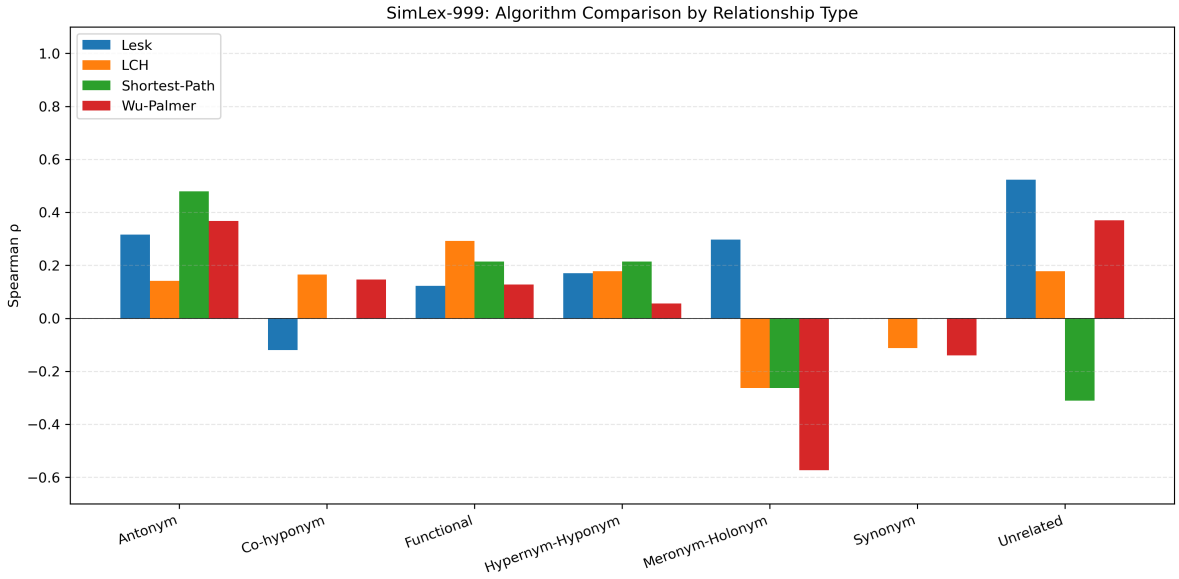


Figure 7: Grouped bar chart comparing algorithm performance across relationship types on SimLex-999. Bars below zero indicate negative correlations.

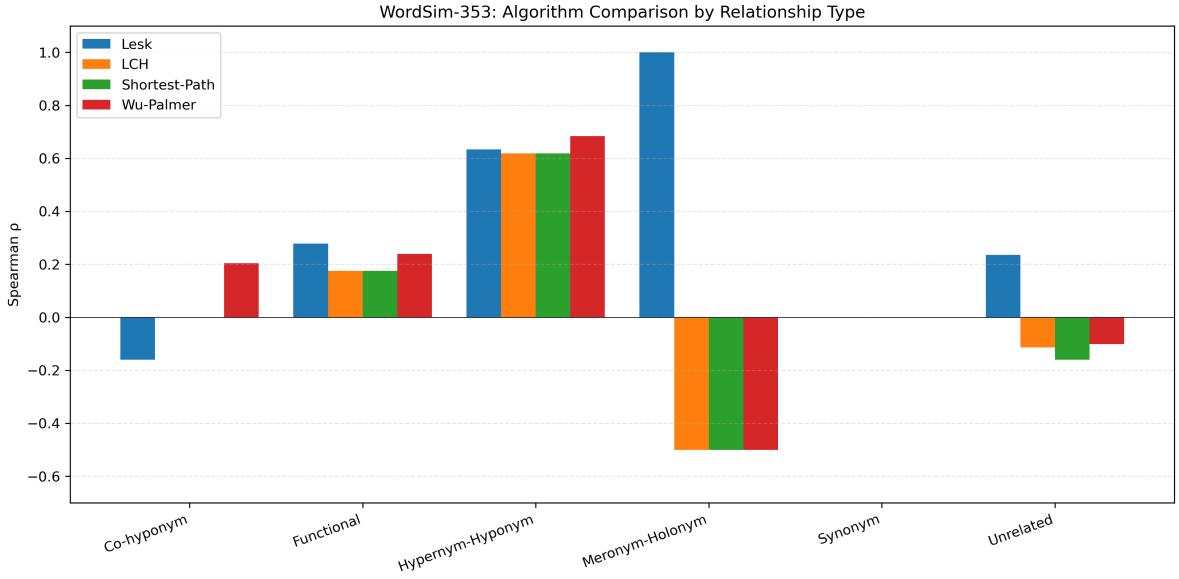


Figure 8: Grouped bar chart comparing algorithm performance across relationship types on WordSim-353.

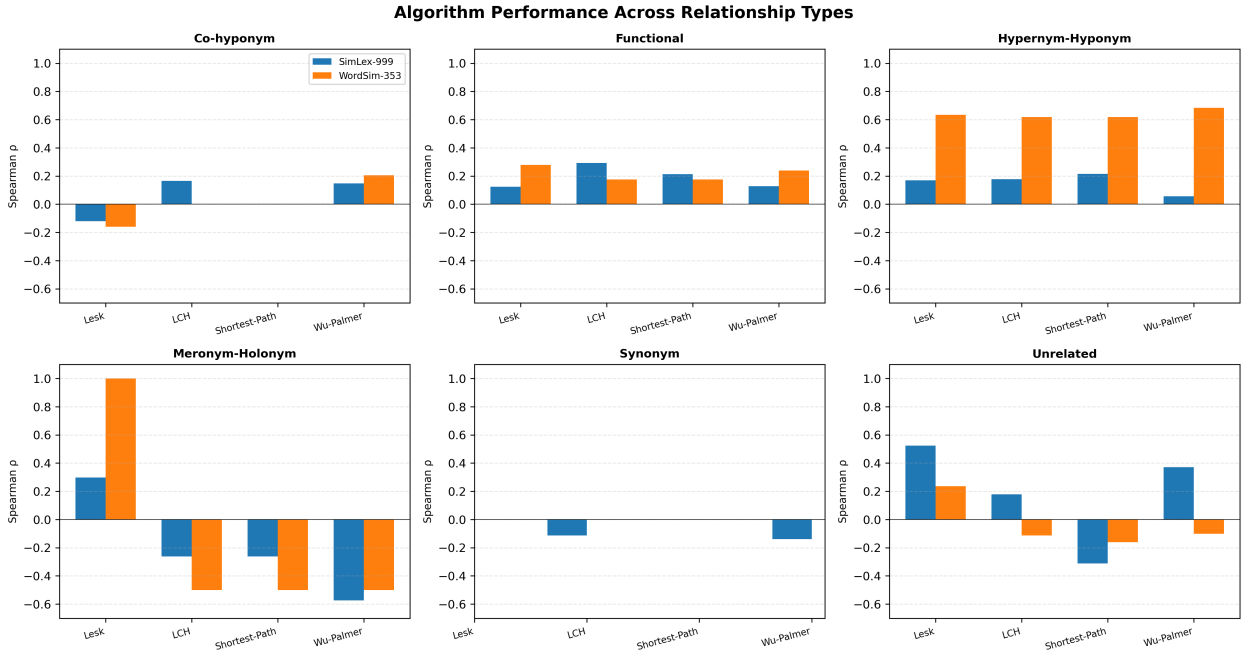


Figure 9: Faceted comparison of algorithm performance on common relationship types across both datasets. Each panel shows SimLex-999 (blue) and WordSim-353 (orange) side-by-side for a specific relationship category.

The heatmaps clearly reveal the complementary strengths: path-based measures (LCH, Shortest-Path, WUP) excel on hypernym–hyponym structures but struggle with meronym/holonym and unrelated pairs, while Lesk achieves its best performance on unrelated pairs due to its sensitivity to gloss dissimilarity. The faceted comparison highlights how hypernym–hyponym pairs are consistently well-handled across both datasets, whereas synonyms and co-hyponyms exhibit near-constant predictions in most algorithms, reducing rank discrimination.

## 3.6 Overall Comparative Analysis of Similarity Algorithms

A comparative evaluation was conducted on four WordNet-based semantic similarity measures: Extended Lesk, Leacock–Chodorow (LCH), Shortest-Path, and Wu–Palmer (WUP)—using two benchmarks with contrasting objectives: SimLex-999, which emphasises strict semantic similarity, and WordSim-353, which rewards broader semantic relatedness. The divergence in algorithmic behaviour across these datasets highlights the fundamental distinction between taxonomic similarity and usage-based relatedness.

### Similarity versus Relatedness Sensitivity

LCH emerges as the strongest performer on SimLex-999, achieving the highest rank correlation among all methods. This confirms that logarithmically scaled taxonomic distance aligns well with human judgments when similarity is defined in terms of conceptual subsumption and shared categorical identity. However, its performance drops on WordSim-353, indicating that path-based hierarchy alone is insufficient to capture the functional, topical, and associative links that dominate relatedness-oriented benchmarks.

Extended Lesk displays the opposite profile. Its strongest performance is observed on WordSim-353, where overlap between extended glosses successfully captures contextual and topical associations that are invisible to hierarchy-based methods. On SimLex-999, however, Lesk underperforms relative to LCH because definitional overlap is a noisy proxy for fine-grained synonymy: distinct concepts may share descriptive vocabulary even when they are not substitutable in usage.

Shortest-Path and Wu–Palmer occupy an intermediate position. Both encode pure structural proximity in the WordNet graph and therefore behave consistently across datasets. Their moderate performance on SimLex confirms that raw and depth-normalised path length capture some component of semantic similarity, but their gains on WordSim remain limited because neither model has access to co-occurrence or functional knowledge beyond the taxonomy.

#### Coverage versus Structural Constraints

Coverage differences among the algorithms reflect architectural limitations of WordNet rather than mere implementation issues. LCH suffers from reduced coverage because it requires a valid connected path within a single part-of-speech hierarchy. All adjective pairs and a nontrivial fraction of verb pairs are therefore discarded, reducing the effective evaluation set. This selective filtering partially explains LCH's superior SimLex correlation: it achieves high accuracy on a restricted subset while ignoring structurally incompatible cases.

In contrast, Extended Lesk, Shortest-Path, and WUP provide near-complete coverage. Lesk is inherently part-of-speech agnostic, as all synsets possess glosses. The path-based methods default to fallback strategies when canonical hierarchical paths are unavailable. This robustness improves aggregate coverage but also admits a class of structurally meaningless similarity estimates that weaken rank fidelity on strict similarity benchmarks.

Overall, the global comparison reveals a clear trade-off: algorithms optimised for taxonomic precision achieve higher peak correlations but sacrifice coverage and robustness, while broader methods offer full coverage at the cost of structural noise.

### 3.7 Performance Analysis by Semantic Relationship Type

Aggregate correlations conceal systematic algorithmic biases that become visible only when performance is conditioned on the underlying semantic relation. We therefore analyse each method across manually annotated relationship categories.

#### Hypernym–Hyponym Relations

All three path-based algorithms perform best on hypernym–hyponym relations in WordSim-353. This is expected, as WordNet was explicitly constructed to optimise vertical IS–A reasoning. Wu–Palmer achieves the strongest performance in this category because its depth normalisation rewards deeply embedded common ancestors more than shallow ones.

On SimLex-999, however, the same relations yield weak correlations. This reflects a fundamental psychological mismatch: humans do not treat hierarchical inclusion as strong semantic similarity. Taxonomic metrics therefore systematically overestimate the similarity of hypernyms relative to human judgments.

#### Meronym–Holonym and Functional Relations

Part–whole relations expose the sharpest failure mode of edge-based similarity. Because meronymy and holonymy are encoded as lateral pointers rather than hierarchical edges in WordNet, all path-

based measures interpret these relations as structurally distant. This results in consistent negative correlations despite strong human judgments of relatedness.

Extended Lesk is the only method that remains stable on meronymy because shared definitional vocabulary directly boosts its similarity estimates. Functional relations exhibit weak but positive correlations across all models. These relations benefit partially from shared categorical structure and partially from gloss co-occurrence, but none of the algorithms is explicitly designed to represent functional dependency.

#### Synonyms and Co-hyponyms

Across both datasets, synonym relations collapse into near-constant predictions for all four algorithms. This is a direct consequence of WordNet's representation: true synonyms are either merged into the same synset or positioned as immediate siblings with identical path distances. As a result, rank correlation becomes undefined or weak despite high absolute similarity.

Co-hyponyms expose a similar structural blind spot. Although such pairs are often perceived as highly similar by humans, path-based measures systematically underestimate their similarity because distance is computed via their parent node rather than through shared functional or distributional properties. Extended Lesk also struggles here because sibling glosses frequently exhibit heavy lexical overlap, again leading to saturation and poor rank discrimination.

#### Antonyms and Unrelated Pairs

Antonyms yield deceptively positive correlations under Shortest-Path and WUP. This reflects structural coincidence rather than semantic success: antonyms frequently share immediate hypernyms and therefore remain close in the graph despite being maximally contrastive in meaning.

Unrelated pairs constitute the only category where Extended Lesk is consistently dominant. Its sensitivity to the absence of shared definitional vocabulary allows it to correctly assign uniformly low similarity to genuinely dissimilar concepts. In contrast, path-based measures often locate unrelated words within moderately shallow tree distances due to WordNet's dense upper ontology, producing spurious similarity.

#### Systematic Error Structure

Across all algorithms, the dominant global error pattern consists of underestimating synonymy and overestimating taxonomic relatedness. This confirms a fundamental representational mismatch between WordNet's graph geometry and human similarity perception: WordNet was designed primarily as a lexical ontology for conceptual classification, not as a perceptual similarity space reflecting interchangeability in usage.

## 4 Discussion

Following are the key observations that tie back to the research questions:

- **Dataset sensitivity (RQ1)**: All measures obtain higher Pearson correlations on WordSim-353 than on SimLex-999, indicating that the WordNet-based scores are more naturally aligned with broad semantic *relatedness* than with the stricter notion of genuine similarity emphasized by SimLex. Among the four, LCH gives the strongest rank correlation on SimLex, while Lesk and WUP are comparatively more competitive on WordSim, where associative and topical links are rewarded.

- **Coverage vs. accuracy trade-off (RQ1)**: LCH achieves the best overall correlations on SimLex but at the cost of reduced coverage: it cannot score adjective pairs and misses some verb pairs when no path exists in the taxonomy, effectively discarding around 19% of the dataset. Lesk, shortest-path, and WUP provide near-complete or full coverage across POS categories, but their higher coverage is offset by lower rank fidelity on the strict similarity benchmark, underscoring a practical trade-off between robustness and peak accuracy.
- **Relationship-specific behavior (RQ2)**: The relationship-type analysis confirms that hypernym/hyponym pairs are consistently the easiest for all three path-based measures, with LCH and WUP in particular showing strong positive correlations on these taxonomic links. In contrast, antonyms, meronym/holonym (part–whole) relations, and many functional associations remain challenging because they are not encoded as short IS–A paths. Lesk, which relies on gloss overlap rather than tree distance, is more tolerant of functional and topical associations but tends to saturate on tightly clustered synonym and co-hyponym sets, producing less nuanced rankings.
- **Limitations of context-free scoring (RQ3)**: All reported results are obtained under a max-over-synsets strategy without true contextual disambiguation. This sense-agnostic approach can artificially inflate similarity for polysemous words by always selecting the pair of senses with the highest possible score, even when those senses are implausible in real usage.

# 5 Conclusion and Next Steps

This study set out to compare four classic WordNet-based measures—Shortest-Path, Leacock–Chodorow (LCH), Wu–Palmer (WUP), and Extended Lesk—along three axes: overall correlation with human similarity/relatedness judgements (RQ1), behavior on specific lexical relationship types (RQ2), and the role of word sense selection (RQ3). The empirical results confirm that no single metric dominates across both SimLex-999 and WordSim-353. LCH achieves the strongest rank alignment on the strict similarity benchmark when its taxonomic assumptions are satisfied, but at the expense of coverage, especially for adjectives and disconnected verbs. Extended Lesk delivers broad, near-POS-agnostic coverage and performs competitively on relatedness-oriented material, while Shortest-Path and WUP provide simple, fully covered baselines that capture much of the same hierarchical information.

The relationship-type analysis clarifies these trade-offs: path-based measures are particularly effective on hypernym–hyponym and closely related taxonomic pairs, yet struggle with antonyms, meronym/holonym relations, and many functional or topical associations that fall outside the IS–A backbone. In contrast, gloss-based Lesk better tolerates loosely associated and topic-related concepts, but tends to saturate on tightly clustered synonym and co-hyponym sets, reducing its ability to finely rank near-synonymous items.

A key limitation of the present work with respect to RQ3 is that all reported scores are obtained under a context-free, max-over-synsets strategy; a full context-aware evaluation using sentence-level disambiguation (e.g., via SCWS or a carefully adapted WiC-style benchmark) has not yet been carried out. As future work, we plan to (i) integrate a genuine contextual similarity dataset to quantify how context-aware Lesk and sense selection affect performance on polysemous words, and (ii) extend the comparison to distributional and contextual embeddings (e.g., Word2Vec, GloVe, or transformer-based models), thereby addressing the remaining aspects of RQ3 and the optional RQ4.