Human and Computational Measurement of Semantic Relations

Nash Whaley

10/10/2025

Contents

Background & Related Work

Lexical Semantic Change (LSC)

- The process by which words gain and lose senses over time
- Likely universal across languages
- **Polysemy** is the synchronic, observable result of lexical semantic change [11, 13]

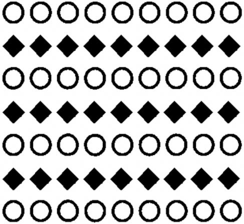
Example

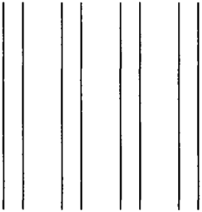
The word *mouse* originally referred to the rodent, but in the 20th century gained a new sense as a computer input device.

Blank's Theory

- 3 semantic relations underlying **association** between concepts: **similarity**, **contiguity**, **contrast** [2]
- Concepts sharing such relations are more strongly associated than those which do not [11, 16].

Psychological Justification





Process of Semantic Change

- When a concept lacks a suitable term, speakers tend to select an existing word for a related concept.
- **Semantic innovation**: A speaker uses a word in a novel way.
- **Lexicalization**: A usage moves from being a discourse norm in an isolated environment and spreads to general usage.

Examples of Lexical Semantic Change

mouse
Context 1: I am shopping for a new **mouse** for my computer.
Context 2: The **mouse** scurried across the kitchen floor.

Eisenbahn
Context 1: Die vorzüglichsten Wägen haben sich im Jahre 1829 auf den **Eisenbahnen** in der Nähe von Glasgow vorgefunden; das Gewicht eines Wagens betrug nämlich ...
Context 2: ... indem bei Ankunft der **Eisenbahn** jederzeit ein Beauftragter von diesem Gasthofe gegenwärtig ist ...

cleave
Context 1: The lumberjack managed to **cleave** the log into two neat pieces.
Context 2: Let my tongue **cleave** to the roof of my mouth.

Types of Innovative LSC

Blank's typology of innovative change includes the following main types [3]:

- **Metaphoric Change** is based on **similarity** between the old and new concepts (e.g., *mouse*, *cloud*)
- **Metonymic Change** arises from **contiguity** between concepts (e.g., Eisenbahn, press).
- **Contrastive change** involves a relation of **contrast** or oppositeness between concepts (cleave, мал [Nenets]).

Although recent progress has been made in LSC detection, existing models fail to differentiate between distinct types of semantic change.

Task Description

Task Definition

arm

Context 1: ... for the conveyance of Mails and Passengers across an **arm** of the sea on the most important route ...

Context 2: Harold was asleep, his bare **arm** thrown above his head, and his eager face relaxed in peace.

4: Completely Applicable
3: Highly Applicable
2: Somewhat Applicable
1: Not Applicable
-: Can't Decide

Annotation Study Organization

- 5 Annotators:
  - Native speakers of North American English
  - All had at least a Bachelor's degree
  - 4 of 5 have had linguistic training
  - Ages 24-32
- 1 annotation for each use-pair per relation
- Guidelines provided before each round
- Successful completion of tutorial before each round

Data Overview

Lemma & Use Selection

- **Canonical lemmas** selected from cognitive semantics literature ([2], [7]), and relevant Wikipedia entries.
- Non-canonical lemmas chosen heuristically from WordNet [5], FrameNet [6], and word-in-context datasets [15].
- Uses selected from the Corpus of Historical American English [1], Merriam-Webster Online [9], and the British National Corpus [4].

Final Dataset Qualities

- 91 lemmas (balanced for predicted strength of relation)
- 631 instances (use-pairs)
- All instances annotated on 4-point scale for each relation
- Gold labels: median annotator label for each instance judgment per task.

CoMeDi Data

CoMeDi dataset [12] used to validate model architecture:

- Multi-lingual semantic relatedness judgments
- Ordinal scale
- Do not distinguish relation types
- Initial fine-tuning on CoMeDi to compensate for sparsity of dataset.

Annotation Results

Are Relations Well-Defined?

- **Inter-Annotator Agreement**: Spearman correlation and Krippendorff's alpha.
- High agreement suggests that some relations may be well-defined
- Aligns broadly with previous LSC studies

Relation Type	**Krippendorff's** α
Similarity	0.59
Contiguity	0.59
Contrast	0.26

Inter-Annotator Agreement

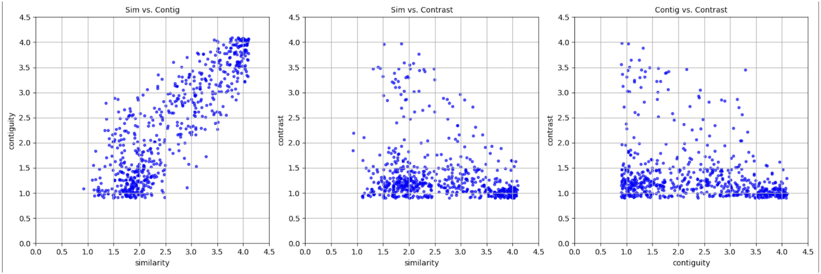
Similarity Spr.

	0	1	2	3	4
0	-	0.543	0.607	0.601	0.642
1	-	-	0.597	0.634	0.614
2	-	-	-	0.651	0.624
3	-	-	-	-	0.651
4	-	-	-	-	-
avg	0.674	0.674	0.723	0.753	0.724

Contiguity Spr.

	0	1	2	3	4
0	-	0.626	0.628	0.543	0.628
1	-	-	0.656	0.649	0.649
2	-	-	-	0.602	0.685
3	-	-	-	-	0.622
4	-	-	-	-	-
avg	0.704	0.765	0.764	0.707	0.741

Are Relations Distinct?



	similarity	contiguity	contrast	sem_rel
similarity	-	0.830	-0.309	0.891
contiguity	-	-	-0.323	0.906
contrast	-	-	-	-0.335
sem_rel	-	-	-	-

Examples

mouth
Context 1: ... headlands on either side of the **mouth** of the harbour could be plainly seen.
Context 2: ... water from the **mouth** of one of the stone lions.
Similarity: 2
Contiguity: 1

gun
Context 1: No, no. He wasn't a bounty hunter. He was a **gun** for hire.
Context 2: I had a run-in with a kid one time and I pulled a weapon on him, I pulled a **gun** on him.
Similarity: 1.4
Contiguity: 2.8

Contrast Task: Canonical vs. Non-Canonical

Non-canonical Lemmas – Spearman Correlation

	0	1	2	3	4
0	-	-0.190902	0.058022	0.101127	0.504461
1	-	-	-0.134293	0.22109	0.062886
2	-	-	-	0.180013	-0.026478
3	-	-	-	-	0.152564
4	-	-	-	-	-
avg	0.383212	0.002352	-0.009465	0.323252	0.339522

Canonical and Non-canonical – Spearman Correlation

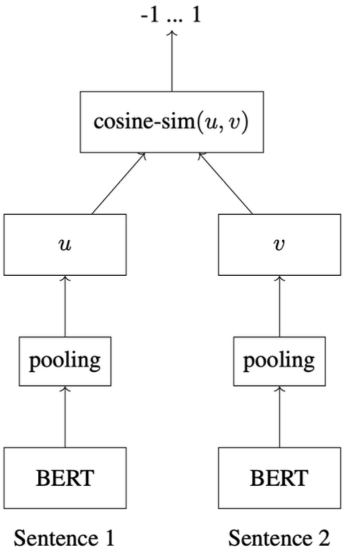
	0	1	2	3	4
0	-	0.618854	0.429477	0.568951	0.366978
1	-	-	0.498652	0.681285	0.279096
2	-	-	-	0.500001	-0.076088
3	-	-	-	-	0.168079
4	-	-	-	-	-
avg	0.642584	0.6389	0.418105	0.605398	0.24556

Model Architecture

Model Architecture

- XLM-R Base Embedding Model
- Max Pooling
- Threshold model maps cosine similarities to ordinal labels
- Eval: Krippendorff (interval) & Spearman

Model Architecture



Experiments

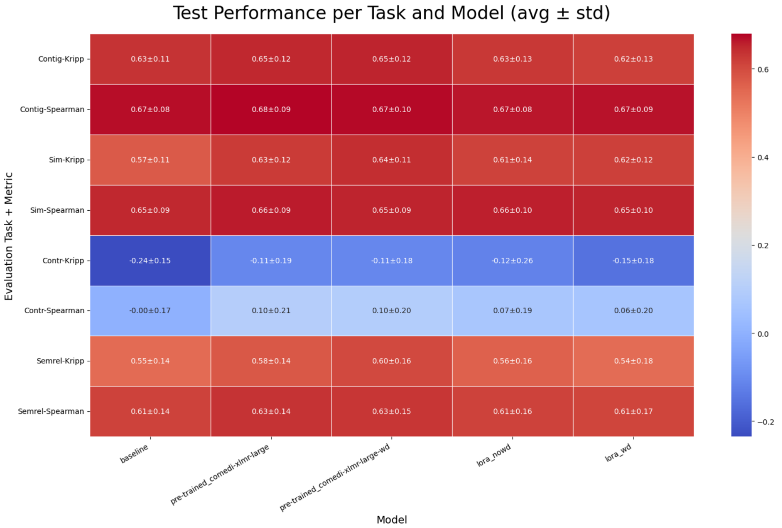
Experimental Setup: Fine-tuning

Hyper-parameters: K-fold Cross-validation, AnglE Loss, Max Seq: 128, Batch Size: 32, early stopping, Train eval: Cosine Similarity + Spr Correlation

- Validation: Models validated on CoMeDi data
- Models trained first on semantic relations datasets only
- Models pretrained on CoMeDi, then fine-tuned for semantic relations

Results

Best Models


Test Performance per Task and Model (avg ± std)
Evaluation Task + Metric
Model
baseline
pre-trained_comedi-xlmr-large
pre-trained_comedi-xlmr-large-wd
lora_nowd
lora_wd
Contig-Kripp 0.63±0.11 0.65±0.12 0.65±0.12 0.63±0.13 0.62±0.13
Contig-Spearman 0.67±0.08 0.68±0.09 0.67±0.10 0.67±0.08 0.67±0.09
Sim-Kripp 0.57±0.11 0.63±0.12 0.64±0.11 0.61±0.14 0.62±0.12
Sim-Spearman 0.65±0.09 0.66±0.09 0.65±0.09 0.66±0.10 0.65±0.10
Contr-Kripp -0.24±0.15 -0.11±0.19 -0.11±0.18 -0.12±0.26 -0.15±0.18
Contr-Spearman -0.00±0.17 0.10±0.21 0.10±0.20 0.07±0.19 0.06±0.20
Semrel-Kripp 0.55±0.14 0.58±0.14 0.60±0.16 0.56±0.16 0.54±0.18
Semrel-Spearman 0.61±0.14 0.63±0.14 0.63±0.15 0.61±0.16 0.61±0.17
0.6
0.4
0.2
0.0
−0.2

Human Performance

Relation Type	Average Performance
Similarity	0.71
Contiguity	0.74
Contrast	0.32

Average annotator agreement with aggregated labels of all other annotators.

Discussion

Discussion

- **Annotation Study:** Strong evidence that **similarity** and **contiguity** are **well-defined and differentiable**.
- Contrast is less reliable, but shows some promise for canonical lemmas.
- **Computational Modeling:** Current models cannot yet distinguish these relations at a human level.
- Higher quality annotations and a larger dataset needed.
- Multimodal approaches might improve model performance.

References I

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde.
CCOHA: Clean Corpus of Historical American English.
In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France, may 2020. European Language Resources Association.

Andreas Blank.
Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen.
Niemeyer, Tübingen, 1997.

Andreas Blank.
Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change.
Historical Semantics and Cognition, pages 61–90, 1999.

References II

BNC Consortium.
British national corpus, xml edition, 2007.
Accessed: 2025-08-15.

Christiane Fellbaum.
Wordnet and wordnets. encyclopedia of language and linguistics, 2005.

Charles J. Fillmore and Beryl T. S. Atkins.
Describing Polysemy: The Case of ‘Crawl’.
In Yael Ravin and Claudia Leacock, editors, *Polysemy: Theoretical and Computational Approaches*. Oxford University Press, 2000.

Dirk Geeraerts.
Prototype theory and diachronic semantics. a case study.
Indogermanische Forschungen (1983), 88(1983):1–32, 1983.

References III

Xianming Li and Jing Li.
Angle-optimized text embeddings, 2024.

Merriam-Webster.
Merriam-webster online dictionary, n.d.
Accessed: 2025-08-15.

Mohammad Taher Pilehvar and Jose Camacho-Collados.
WiC: the word-in-context dataset for evaluating context-sensitive meaning representations.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

References IV

Dominik Schlechtweg.
Human and Computational Measurement of Lexical Semantic Change.
Stuttgart, Germany, 2023.

Dominik Schlechtweg, Tejaswi Choppa, Wei Zhao, and Michael Roth.
CoMeDi shared task: Median judgment classification & mean disagreement ranking with ordinal word-in-context judgments.
In Michael Roth and Dominik Schlechtweg, editors, *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 33–47, Abu Dhabi, UAE, jan 2025. International Committee on Computational Linguistics.

References V

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi.
SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection.
In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain, 2020. Association for Computational Linguistics.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann.
Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 169–174, New Orleans, Louisiana, 2018.

References VI

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray.
DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages.
In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic, nov 2021. Association for Computational Linguistics.

Sachin Yadav and Dominik Schlechtweg.
XL-DURel: Finetuning sentence transformers for ordinal Word-in-Context classification, 2025.

Measuring LSC

The standard annotation approach:

- Use-pair annotation
- Four-point Scale
- DURel [14] scale adapted from Blank's theory [11, 33].

4: Identical	Identity
3: Closely Related	Context Variance
2: Distantly Related	Polysemy
1: Unrelated	Homonymy

Measuring LSC

- Use-pair annotations can be represented as graph
  - Nodes correspond to word uses
  - Edges are weighted by aggregated relatedness scores
- Time-specific allow comparison of meanings across periods. Clustering algorithms use edge weights to group unique word senses.
- **Binary** change measured by loss or gain of cluster in the time dimension.
- **Graded** change (introduced in SemEval shared task [13]) measured as a continuous value between 0 and 1 (e.g., cosine). Benchmarks are currently available in many languages.

Computational Models of LSC

- Computational models of meaning typically fall into two categories:
  - **Token-based models**: create distinct representations for each *use* of a word (context-sensitive).
  - **Type-based models**: aggregate over all occurrences to form a single, generalized representation.

Token-based models, e.g. Fine-tuned **Word-in-Context** (WiC) models [10], use contextual embeddings from **BERT** or **XLM-R**. These have been shown to outperform type-based models.

WiC Models

Typical WiC setup:

- Input: A pair of uses of the same target word.
- Contextual embedding model: produces embeddings for each occurrence.
- Vector processor: aggregates embeddings (e.g., concatenation)
- Classifier head: predicts semantic relatedness (often cosine similarity). A threshold classifier may be used to make binary predictions.

Current SotA models improve on previous WiC models by fine-tuning the contextual embedding model.

Ordinal models of LSC

- **OGWiC (Ordinal Graded Word-in-Context)** [12] extends the traditional WiC task by introducing **ordinal prediction** on the four-point DURel scale.
  - Moves beyond binary classification to capture **graded semantic similarity**.
  - Enables direct comparison with **human ordinal judgments**.
  - Evaluated using **Krippendorff's** α to assess agreement with annotators.
- **XL-DURel** (SotA) [16]:
  - Based on **Sentence-BERT** with **XLM-R-Large** as the base embedder.
  - Uses **max pooling** to derive full-sentence representations.
  - Optimized for ordinal similarity through **AnglE loss** [8], a ranking-based loss emphasizing **angular distances** in embedding space.
  - Better captures **graded semantic similarity** between word uses.

Canonical Lemmas

Canonical Lemmas (per relation type):

- **Similarity:** `see`, `lure`, `artillery`, `arm`
- **Contiguity:** `canine`, `sweat`, `tongue`, `press`
- **Contrast:** `bad`, `consult`, `dust`, `sanction`

Similarity Task: Canonical vs Non-Canonical Spearman

Non-Canonical

	0	1	2	3	4
0	-	0.440168	0.457323	0.622887	0.529834
1	-	-	0.547888	0.575912	0.625609
2	-	-	-	0.787721	0.659627
3	-	-	-	-	0.733422
4	-	-	-	-	-
avg	0.575586	0.668814	0.739646	0.818149	0.772523

Canonical and Non-canonical

	0	1	2	3	4
0	-	0.602989	0.603598	0.7363	0.625053
1	-	-	0.568235	0.674241	0.646902
2	-	-	-	0.810168	0.682609
3	-	-	-	-	0.755977
4	-	-	-	-	-
avg	0.705657	0.675382	0.738938	0.860779	0.76885

Contiguity Task: Canonical vs. Non-Canonical Spearman

Non-canonical Lemmas – Spearman Correlation

	0	1	2	3	4
0	-	0.707369	0.712097	0.577492	0.568273
1	-	-	0.711139	0.655335	0.742818
2	-	-	-	0.530008	0.654127
3	-	-	-	-	0.516516
4	-	-	-	-	-
avg	0.753208	0.851197	0.756893	0.666853	0.653441

Canonical and Non-canonical – Spearman Correlation

	0	1	2	3	4
0	-	0.591529	0.528091	0.48575	0.552732
1	-	-	0.587595	0.688872	0.613411
2	-	-	-	0.393474	0.562761
3	-	-	-	-	0.380927
4	-	-	-	-	-
avg	0.651585	0.80508	0.605713	0.590214	0.582991

Contrast Task: Canonical vs. Non-Canonical

Non-canonical Lemmas – Spearman Correlation

	0	1	2	3	4
0	-	-0.190902	0.058022	0.101127	0.504461
1	-	-	-0.134293	0.22109	0.062886
2	-	-	-	0.180013	-0.026478
3	-	-	-	-	0.152564
4	-	-	-	-	-
avg	0.383212	0.002352	-0.009465	0.323252	0.339522

Canonical and Non-canonical – Spearman Correlation

	0	1	2	3	4
0	-	0.618854	0.429477	0.568951	0.366978
1	-	-	0.498652	0.681285	0.279096
2	-	-	-	0.500001	-0.076088
3	-	-	-	-	0.168079
4	-	-	-	-	-
avg	0.642584	0.6389	0.418105	0.605398	0.24556

Inter-Annotator Agreement: Similarity Task

Spearman (avg. leave-one-out)

	0	1	2	3	4
0	-	0.543	0.607	0.601	0.642
1	-	-	0.597	0.634	0.614
2	-	-	-	0.651	0.624
3	-	-	-	-	0.651
4	-	-	-	-	-
avg	0.674	0.674	0.723	0.753	0.724

Inter-Annotator Agreement: Contiguity Task

Spearman (avg. leave-one-out)

	0	1	2	3	4
0	-	0.626	0.628	0.543	0.628
1	-	-	0.656	0.649	0.649
2	-	-	-	0.602	0.685
3	-	-	-	-	0.622
4	-	-	-	-	-
avg	0.704	0.765	0.764	0.707	0.741

Inter-Annotator Agreement: Contrast Task

Spearman (avg. leave-one-out)

	0	1	2	3	4
0	-	0.415	0.256	0.432	0.171
1	-	-	0.378	0.566	0.215
2	-	-	-	0.312	0.104
3	-	-	-	-	0.272
4	-	-	-	-	-
avg	0.348	0.401	0.227	0.475	0.159