Human and Computational Measurement of Semantic Relations

Nash Whaley & Dominik Schlechtweg

28/08/2026

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]

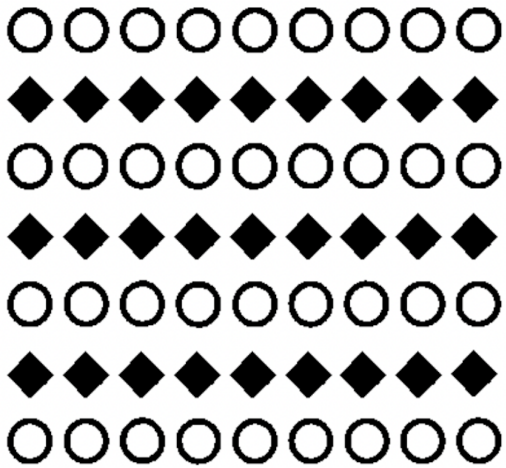
Example

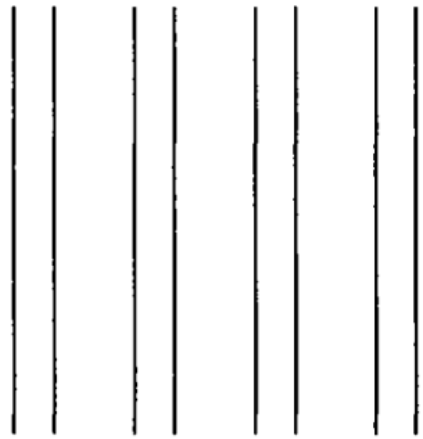
The word *mouse* originally referred to the rodent, but in the 20th century gained a new sense as a computer input device.

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 . . .

bad

Context 1: I don't like you, Mr. Porter. I've tolerated your **bad** manners because your mother and I were friends.
Context 2: he is the funniest man in America, and, after Muhammad Ali, he is the **baddest** person anywhere

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*)
- **Metonymic Change** arises from **contiguity** between concepts (e.g., *Eisenbahn*).
- **Contrastive change** involves a relation of **contrast** or oppositeness between concepts (e.g., *bad*).

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

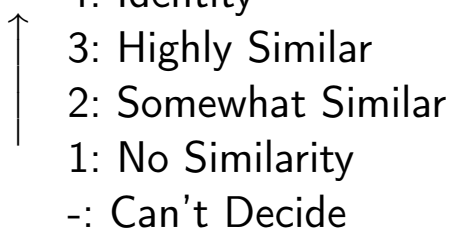
Task Description

Task Definition

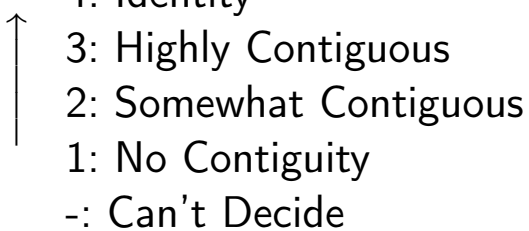
arm

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

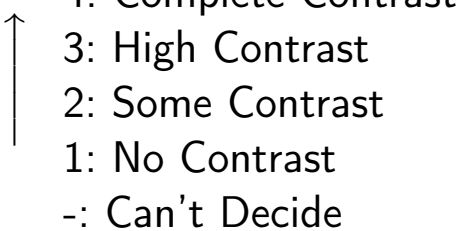
Context 2: Harold was asleep, his bare **arm** thrown above his head, and his eager face relaxed in peace.



Similarity Scale



Contiguity Scale



Contrast Scale

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 extracted 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.
- Dataset published: https://zenodo.org/records/20801951

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.

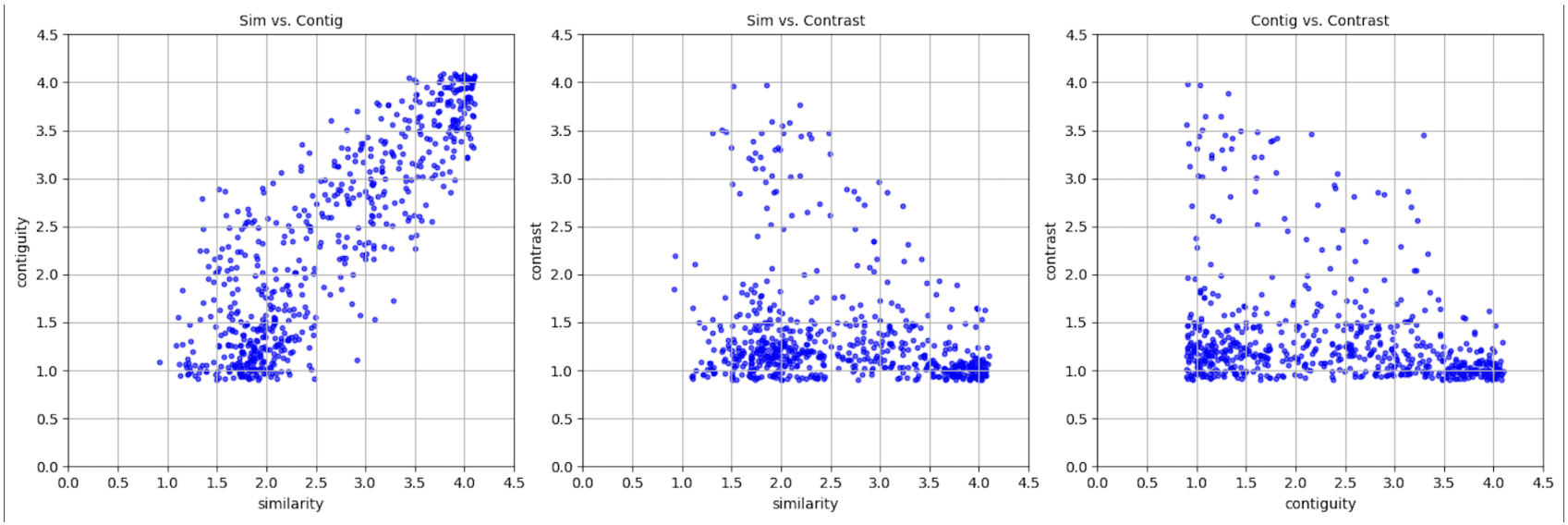
Annotation Results

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

Are Relations Distinct?



	similarity	contiguity	contrast
similarity	-	0.830	-0.309
contiguity	-	-	-0.323
contrast	-	-	-

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

- Random sample of non-canonical judgments
- Combined sample of non-canonical and canonical judgments of same size
- Stronger inter-annotator agreement when canonical examples present.

	Avg. Pairwise Agreement (spr)
Non-canonical Sample	0.09
Combined Sample	0.4

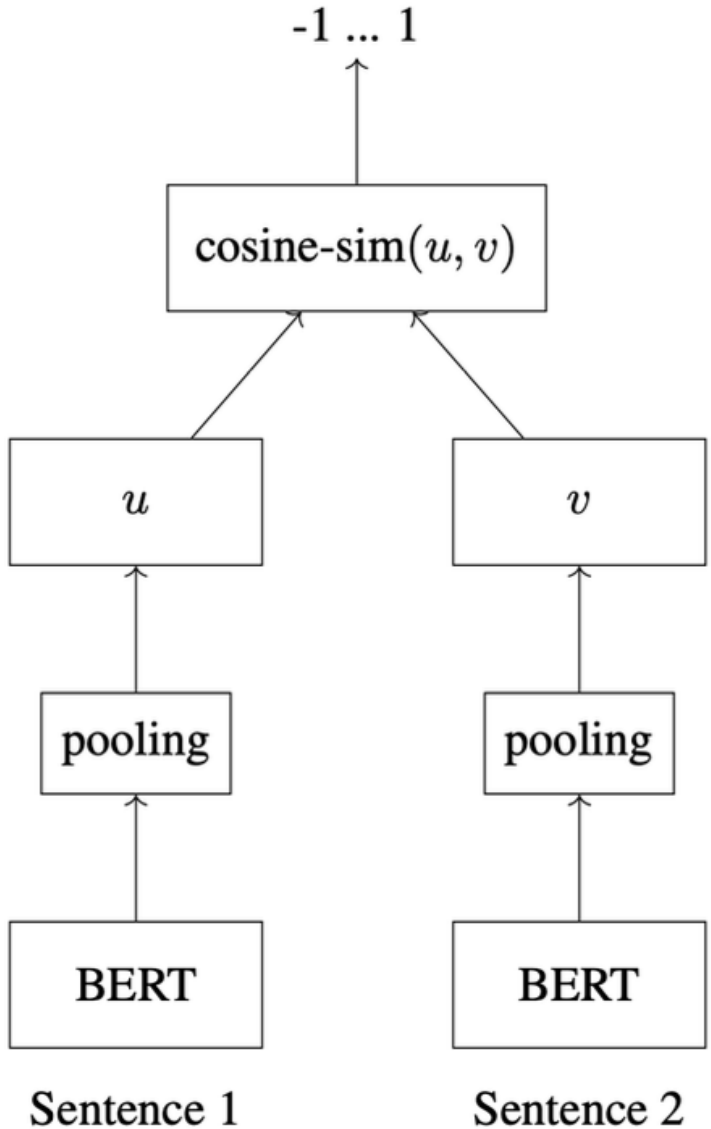
Computational Models

Can we reach human-like annotation performance?

Encoder-based

- Lightweight, adaptable, and transparent
- Commonly used in semantic relatedness studies
- XLM-R Base **Embedding Model** (multi-lingual)
- Threshold model maps **cosine similarities** to **ordinal labels**
- Eval: Krippendorff (interval) & Spearman

Encoder-based



Fine-tuning

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

LLM-based

- Powerful and easy to use
- Model: `GPT-5.3-chat-latest`
- Approach adapted from [17]
- Prompt included examples from tutorial and solicited ordinal label (1-4)

Results

Human Annotator vs. Model Performance

Human Leave-One-Out Alphas

Annotator	Similarity	Contiguity	Contrast
1	0.69	0.59	0.56
2	0.63	0.71	**0.64**
3	0.70	**0.73**	0.54
4	**0.75**	0.7	**0.64**
5	0.72	**0.73**	0.36

Human Annotator vs. Model Performance

Model Leave-One-Out Alphas

Left Out	XLMR Sim	GPT Sim	XLMR Contig	GPT Contig	XLMR Contr.	GPT Contr.
Annotator 1	0.65	0.74	0.67	**0.72**	-0.02	0.63
Annotator 2	0.66	0.73	0.66	0.65	-0.02	0.62
Annotator 3	0.65	0.74	0.66	0.68	-0.03	0.65
Annotator 4	0.67	0.73	0.67	0.7	-0.06	0.55
Annotator 5	0.67	**0.76**	0.67	0.68	-0.04	**0.66**

Human Annotator vs. Model Performance

Leave-One-Out Average Krippendorff's Alpha Scores

Metric	Human Average	GPT Average	XLMR Model Avg
Similarity	0.7	**0.74**	0.66
Contiguity	**0.69**	**0.69**	0.66
Contrast	0.55	**0.62**	-0.03

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:** LLMs appear capable of **human level** annotation in few-shot setting.
- Fine-tuned models approach lower end of human annotation for contiguity and similarity.

Future Work

- Multi-modal approaches might improve model performance at lower computational cost.
- More annotations needed.
- Data could be augmented with metonymy and metaphor datasets.

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.

References VII

Д. В. Кокосинский.
Большие языковые модели в задаче оценки семантической близости словоупотреблений.
Электронные библиотеки, 29(4):1133–1154, july 2026.

Measuring LSC

The standard annotation approach:

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

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

Similarity Prompt

You are a highly trained text data annotation tool capable of providing subjective responses. Rate the semantic similarity of the target word in each sentence pair below. Consider only the objects/concepts the word forms refer to: ignore any common etymology! Ignore case! Ignore number (cat/Cats = identical meaning). Homonyms (like bat the animal vs bat in baseball) count as unrelated. Output numeric rating: 1 is no similarity; 2 is somewhat similar; 3 is very similar; 4 is identical meaning. Your response should align with a human's succinct judgment. You are annotating for similarity which means that the concepts denoted by the target word in each sentence must share some overlapping features (visually or conceptually).
Example: target word: bank first sentence: His parents had left a lot of money in the bank and now it was all Measle's, but a judge had said the Measle was too young to get it. second sentence: Sherrell, it is said, was sitting on the bank of the river close by, and as soon as the men had disappeared from sight he jumped on board the schooner. Score: 1
Example: target word: tree first sentence: A binary tree is a data structure for storing organized data. second sentence: The trees in that forest are rich in aroma and colors. Score: 2
Example: target word: cell first sentence: With this well-known program, we will be able to manipulate the data through its cells. second sentence: As a result of this analysis, we will obtain a different map, a map where each visible cell comes with a number of 0 or 180 degrees. Score: 3
Example: target word: horse first sentence: Marti's white horse gallops to victory. second sentence: When horses are in a herd, they protect each other. Score: 4
Now score each of the following numbered items. For each item, output ONLY a line in the exact format: N: score where N is the item number and score is 1, 2, 3, or 4. Do not explain your reasoning. Do not add any other text. Output exactly one line per item, in order.

Contiguity Prompt

You are a highly trained text data annotation tool capable of providing subjective responses. Rate the semantic contiguity of the target word in each sentence pair below. Consider only the objects/concepts the word forms refer to: ignore any common etymology! Ignore case! Ignore number (cat/Cats = identical meaning). Homonyms (like bat the animal vs bat in baseball) count as unrelated. Output numeric rating: 1 is no contiguity; 2 is some contiguity; 3 is high contiguity (often strong example of metonymy); 4 is identical meaning. A contiguity relationship occurs when two concepts denoted by a word belong to the same field of knowledge and are spatially, temporally, or causally connected (i.e., frequently experienced together). You will be shown a pair of sentences containing the target word. Your response should align with a human's succinct judgment. You are annotating for contiguity which means that the concepts denoted by the target word in each sentence must share some semantic frame.
Example: target word: bank first sentence: His parents had left a lot of money in the bank and now it was all Measle's, but a judge had said the Measle was too young to get it. second sentence: Sherrell, it is said, was sitting on the bank of the river close by, and as soon as the men had disappeared from sight he jumped on board the schooner. Score: 1
Example: target word: blink first sentence: Today, art co-opted by advertising is so commonplace we do not blink at Michelangelo's David wearing Levi's. second sentence: When I glance back to the coral, the ocean is transformed. I blink, thinking I'm seeing things. Score: 2
Example: target word: body first sentence: We could use some strong bodies for our team. second sentence: I try to exercise as much as I can to keep my body in shape. Score: 3
Example: target word: horse first sentence: Marti's white horse gallops to victory. second sentence: When horses are in a herd, they protect each other. Score: 4
Now score each of the following numbered items. For each item, output ONLY a line in the exact format: N: score where N is the item number and score is 1, 2, 3, or 4. Do not explain your reasoning. Do not add any other text. Output exactly one line per item, in order.

Contrast Prompt

You are a highly trained text data annotation tool capable of providing subjective responses. Rate the semantic contrast of the target word in each sentence pair below. Consider only the objects/concepts the word forms refer to: ignore any common etymology! Ignore case! Ignore number (cat/Cats = identical meaning). Homonyms (like bat the animal vs bat in baseball) count as unrelated (i.e., no contrast) because they cannot be meaningfully be said to be opposite each other. Where there is no contrast between the concepts conveyed in each use pair, uses are either identical, completely unrelated, or related in a way that does not involve contrast. Output numeric rating: 1 is no contrast; 2 is some contrast; 3 is high contrast; 4 is diametrically opposite meanings. A contrast relationship occurs when two concepts denoted by a word are opposite to some degree from one another in meaning. You will be shown a pair of sentences containing the target word. Your response should align with a human's succinct judgment.
Example: target word: mouse first sentence: The cat chased the mouse under the back porch. second sentence: When I was a child I had a pet mouse. Score: 1
Example: target word: exploit first sentence: The cruel factory owner exploited the workers. second sentence: The athlete exploited her speed. Score: 2
Example: target word: rock first sentence: She rocked the baby in the crib. second sentence: The earthquake rocked the city. Score: 3
Example: target word: cleave first sentence: When the dust groweth into hardness, and the clods cleave fast together? second sentence: It's a piece of cake for me to cleave the brick in two. Score: 4
Now score each of the following numbered items. For each item, output ONLY a line in the exact format: N: score where N is the item number and score is 1, 2, 3, or 4. Do not explain your reasoning. Do not add any other text. Output exactly one line per item, in order.

Semrel Prompt

You are a highly trained text data annotation tool capable of providing subjective responses. Rate the semantic similarity of the target word in each sentence pair below. Consider only the objects/concepts the word forms refer to: ignore any common etymology and metaphorical similarity! Ignore case! Ignore number (cat/Cats = identical meaning). Homonyms (like bat the animal vs bat in baseball) count as unrelated. Output numeric rating: 1 is unrelated; 2 is distantly related; 3 is closely related; 4 is identical meaning. Your response should align with a human's succinct judgment.
Example: target word: bank first sentence: His parents had left a lot of money in the bank and now it was all Measle's, but a judge had said the Measle was too young to get it. second sentence: Sherrell, it is said, was sitting on the bank of the river close by, and as soon as the men had disappeared from sight he jumped on board the schooner. Score: 1
Example: target word: tree first sentence: A binary tree is a data structure for storing organized data. second sentence: The trees in that forest are rich in aroma and colors. Score: 2
Example: target word: cell first sentence: With this well-known program, we will be able to manipulate the data through its cells. second sentence: As a result of this analysis, we will obtain a different map, a map where each visible cell comes with a number of 0 or 180 degrees. Score: 3
Example: target word: horse first sentence: Marti's white horse gallops to victory. second sentence: When horses are in a herd, they protect each other. Score: 4
Now score each of the following numbered items. For each item, output ONLY a line in the exact format: N: score where N is the item number and score is 1, 2, 3, or 4. Do not explain your reasoning. Do not add any other text. Output exactly one line per item, in order.

LOO Models Vs Annotators

Left Out	XLMR Sim	GPT Sim	XLMR Contig	GPT Contig	XLMR Contr.	GPT Contr.
Annotator 1	0.65	0.74	0.67	0.72	-0.02	0.63
Annotator 2	0.66	0.73	0.66	0.65	-0.02	0.62
Annotator 3	0.65	0.74	0.66	0.68	-0.03	0.65
Annotator 4	0.67	0.73	0.67	0.7	-0.06	0.55
Annotator 5	0.67	0.76	0.67	0.68	-0.04	0.66