University of Stuttgart
Germany

UTN Technische Universität Nürnberg

CoMeDi Shared Task: Median Judgment Classification & Mean Disagreement Ranking with Ordinal Word-in-Context Judgments

February 5, 2025

Dominik Schlechtweg, Tejaswi Choppa, Wei Zhao, Michael Roth
University of Stuttgart, University of Aberdeen, University of Technology Nuremberg

IMS

1495 UNIVERSITY OF ABERDEEN

Introduction

- near-human performance in several semantic NLP tasks (A. Wang et al., 2019)
- e.g. WiC (Pilehvar & Camacho-Collados, 2019)
  - asking if the same word in two contexts has the same meaning
  - **binary classification**
  - elegant simplification of classical WSD
  - state-of-art model has obtained near-human performance, 77.9% vs. 80% (Z. Wang et al., 2021)
  - avoids need for sense glosses
  - inadequate simplification

Ordinal Graded Word-in-Context Classification

- more theory-adequate formulation: GWiC (Armendariz et al., 2020)
  - asking to provide graded WiC predictions
  - did not require to **reproduce** human annotations
  - **ranking task**
  - can be fulfilled by predictions on an arbitrary scale
  - exactly reproducing human annotations has advantages such as providing linguistic interpretations
- **Ordinal Graded Word-in-Context Classification** (OGWiC)
  - asking participants to **exactly reproduce** instance labels instead of just inferring their relative order

Disagreement in Word-in-Context Ranking

- WiC datasets annotated on ordinal scales show considerable **disagreement**

  (cf. Schlechtweg et al., 2024)
- traditional aggregation leads to information loss
- modeling disagreement is important for **realistic** scenarios
- predict on items where high disagreement is expected
- can help to detect or filter highly complicated samples
- recent research uses alternative aggregation techniques

  (e.g. Leonardelli et al., 2023; Uma et al., 2022)
- **Disagreement in Word-in-Context Ranking** (DisWiC)
  - asking to predict the **amount** of disagreement
  - differs from previous tasks by making disagreement the explicit ranking aim

Shared Task

- both tasks introduced in **CoMeDi shared task** (Schlechtweg et al., 2025)
- data, starting kits, codalab, results, papers:

`https://comedinlp.github.io/`

Task Definitions

(1) ...and taking a knife from her pocket, she opened a vein in her little **arm**.
(2) ...and though he saw her within reach of his **arm**, yet the light of her eyes seemed as far off.

- Sample judgments: [4,4]; median: 4; mean pairwise difference: 0.0

(1) ...and taking a knife from her pocket, she opened a vein in her little **arm**.
(3) It stood behind a high brick wall, its back windows overlooking an **arm** of the sea which, at low tide, was a black and stinking mud-flat.

- Sample judgments: [2,3,2]; median: 2; mean pairwise difference: 0.667

- **OGWiC**: For each usage pair, predict the **median** of annotator judgments
- **DisWiC**: For each usage pair, predict the **mean of pairwise absolute judgment differences between** annotator judgments

Annotation Scale

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

Table 1: The DURel relatedness scale (Schlechtweg et al., 2018) on the left and its interpretation from Schlechtweg (2023, p. 33) on the right.

Data

- use publicly available **ordinal WiC datasets** from multiple languages:

  `https://www.ims.uni-stuttgart.de/data/wugs`
- provide large number of judgments for **word usage pairs** on the DURel scale
- have so far not been used primarily for WiC-like tasks
- augment with roughly 33k unpublished instances
- ensure data quality through overall agreement and cleaning

Datasets

Dataset	LG	Reference	JUD	VER	KRI	SPR
ChiWUG	ZH	Chen et al. (2023)	61k	1.0.0	.60	.69
DWUG	EN	Schlechtweg et al. (2021)	69K	3.0.0	.63	.55
DWUG Res.	EN	Schlechtweg et al. (2024)	7K	1.0.0	.56	.59
DWUG	DE	Schlechtweg et al. (2021)	63K	3.0.0	.67	.61
DWUG Res.	DE	Schlechtweg et al. (2024)	10K	1.0.0	.59	.7
DiscoWUG	DE	Kurtyigit et al. (2021)	28K	2.0.0	.59	.57
RefWUG	DE	Schlechtweg (2023)	4k	1.1.0	.67	.7
DURel	DE	Schlechtweg et al. (2018)	6k	3.0.0	.54	.59
SURel	DE	Hätty et al. (2019)	5k	3.0.0	.83	.84
NorDiaChange	NO	Kutuzov et al. (2022)	19k	1.0.0	.71	.74
RuSemShift	RU	Rodina and Kutuzov (2020)	8k	1.0.0	.52	.53
RuShiftEval	RU	Kutuzov and Pivovarova (2021)	30k	1.0.0	.56	.55
RuDSI	RU	Aksenova et al. (2022)	6k	1.0.0	.41	.56
DWUG	ES	Zamora-Reina et al. (2022)	62k	4.0.1	.53	.57
DWUG	SV	Schlechtweg et al. (2021)	55K	3.0.0	.67	.62
DWUG Res.	SV	Schlechtweg et al. (2024)	16K	1.0.0	.56	.65

Table 2: Datasets used for our task. All are annotated on the DURel scale.

Cleaning and Aggregation

1. pre-cleaning
2. cleaning
   - exclude instances with less than two judgments
   - exclude instances with "Cannot decide" judgments, strong disagreement and non-integer median (OGWiC)
   - ignore "Cannot decide" judgments (DisWiC)
3. aggregation
4. split
   - per language into train/test/dev (70/20/10%)
   - **at target words** (lexical split)
   - no overlap in target words and uses

Data statistics

Task	# Instances	# Uses	# Lemmas	Split
OGWiC	48K	55K	520	Train
	8K	8K	77	Dev
	15K	16K	152	Test
DisWiC	82K	55K	521	Train
	13K	8K	77	Dev
	26K	16K	152	Test

Table 3: Data statistics after cleaning and aggregation per split and and over all languages combined.

Data Distribution (OGWiC)

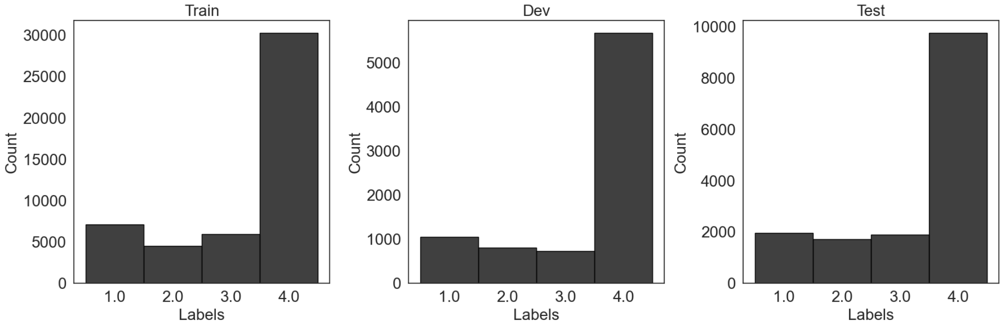

Figure 1: Label distribution for OGWiC task for all languages combined.

Data Distribution (DisWiC)

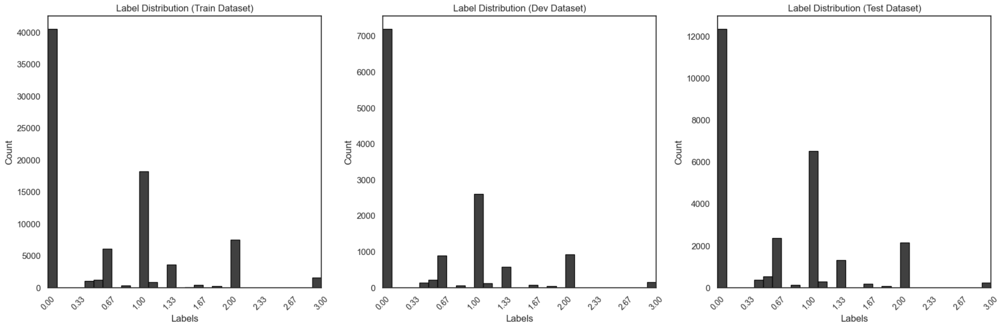

Figure 2: Label distribution for DisWiC task for all languages combined.

Models (Baselines)

- baseline models: (cf. Choppa et al., 2025)
  - **Baseline 1**: XLM-R + CosTH
  - **Baseline 2**: XL-Lexeme + CosTH
  - **Baseline 3**: XLM-R + LR
  - **Baseline 4**: XL-Lexeme + MLP
  - **Upper bound** (OGWiC)
- **structure**:
  1. vectorize usages with contextualized encoder
  2. concatenate vectors
  3. classify with threshold on cosine similarity/linear regression/multi-layer perceptron
- Baselines 2 and 3 were published in development phase

Models (Participants)

- 5 teams participated
  - **Deep-change** (Kuklin & Arefyev, 2025)
  - **GRASP** (Alfter & Appelgren, 2025)
  - **MMLabUIT** (Le & Van, 2025)
  - **JuniperLiu** (Liu et al., 2025)
  - **FuocChu_VIP123** (Chu, 2025)
- often (but not always) similar to baseline models
- three unofficial submissions (Choppa et al., 2025; Loke et al., 2025; Sarumi et al., 2025)

Evaluation

- **OGWiC**: Krippendorff's α (ordinal version) (Krippendorff, 2018)
  - penalizes stronger deviations from the gold label more heavily
  - controls for expected disagreement
  - is recommended for ordinal classification (Sakai, 2021)
- **DisWiC**: Spearman's ρ (Spearman, 1904)
  - measures correspondence of rankings according to amount of disagreement
- in **development** phase, the starting kits, training and development data released
- in **evaluation** phase, public test instances were published and participants were allowed to make **3 submissions**
- leaderboard on Codalab was kept hidden during evaluation phase
- hidden gold labels were published during post-evaluation phase

Timeline

Task	Development Phase	Evaluation Phase
OGWiC	August 23–September 14	October 14–21
DisWiC	September 15–October 13	October 21–27

Table 4: Shared task timeline.

Results

Task	Team	AV	-ES	ZH	EN	DE	NO	RU	ES	SV
OGWiC	Upper bound	.95	.95	1.	.97	.88	.94	.96	.96	.95
	deep-change	**.66**	**.64**	**.42**	**.73**	.72	**.67**	**.62**	**.75**	**.68**
	Baseline 2	.58	.57	.38	.65	**.73**	.52	.55	.66	.60
	GRASP	.56	.54	.32	.56	.66	.59	.49	.64	.65
	MMLabUIT	.52	.51	.36	.57	.67	.44	.42	.60	.61
	JuniperLiu	.27	.26	.14	.51	.49	.08	.13	.33	.22
	Baseline 1	.12	.12	.06	.10	.27	.12	.11	.18	.02

Table 5: Top results of OGWiC evaluation phase. 'AV' = Average over languages; '-ES' = Average over languages excluding Spanish.

Results

Task	Team	AV	-ES	ZH	EN	DE	NO	RU	ES	SV
DisWiC	deep-change	**.23**	.23	.30	**.08**	**.20**	**.29**	**.18**	**.19**	**.35**
	GRASP	.22	.23	**.54**	.04	.11	.27	.17	.12	.30
	Baseline 4	.16	.17	.49	.06	.09	.24	.12	.08	.08
	FuocChu.	.12	.14	.36	.02	.10	.16	.05	.01	.17
	Baseline 3	.12	.12	.39	.06	.09	.08	.05	.08	.08
	JuniperLiu	.08	.09	.36	.04	.02	-.04	.07	.04	.09
	sunfz1	.07	.07	.30	.05	-.00	-.07	.07	.04	.09

Table 6: Top results of DisWiC evaluation phase. ‘AV’ = Average over languages; ‘-ES’ = Average over languages excluding Spanish; ‘FuocChu.’ = FuocChu_VIP123.

Conclusion

- introduced **two new tasks** based on ordinal Word-in-Context annotations between word usages
  - OGWiC
  - DisWiC
- **OGWiC** solved with rather high performance
- **DisWiC** remains a challenge
  - on some languages performance is exceptionally high
- both tasks dominated by same teams employing a **Word-in-Context model optimized on independent binary Word-in-Context data**
- dominant approach to solve OGWiC was **thresholding of graded similarity predictions**

Future Work

- solve the two tasks with **different data splitting** conditions
- tie the published test data to **individual annotators**

Limitations

- influence of **annotator number** on mean disagreement values
  - control number of annotations per instance or provide at test time
  - explore other disagreement measures
- **narrowness** of training, development and test data in terms of target words
  - avoid lexical split
- Krippendorff's α estimates the **expected label distribution** from both model and gold labels
  - explore modifications of Krippendorff's α estimating the expected label distribution solely from the gold data
- performance **upper bound** influenced by random agreement
  - report results for historical and modern language instances separately
- ignorance of **diachronic** dataset component

References I

Aksenova, A., Gavrishina, E., Rykov, E., & Kutuzov, A. (2022, October). RuDSI: Graph-based word sense induction dataset for Russian. In D. Ustalov et al. (Eds.), *Proceedings of textgraphs-16: Graph-based methods for natural language processing* (pp. 77–88). Gyeongju, Republic of Korea: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.textgraphs-1.9

Alfter, D., & Appelgren, M. (2025). GRASP at CoMeDi Shared Task: Multi-strategy modeling of annotator behavior in multi-lingual semantic judgments. In *Proceedings of the 1st workshop on context and meaning–navigating disagreements in nlp annotations.* Abu Dhabi, UAE.

Armendariz, C. S., Purver, M., Pollak, S., Ljubešić, N., Ulčar, M., Vulić, I., & Pilehvar, M. T. (2020, December). SemEval-2020 task 3: Graded word similarity in context. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (Eds.), *Proceedings of the fourteenth workshop on semantic evaluation* (pp. 36–49). Barcelona (online): International Committee for Computational Linguistics. Retrieved from https://aclanthology.org/2020.semeval-1.3 doi: 10.18653/v1/2020.semeval-1.3

Chen, J., Chersoni, E., Schlechtweg, D., Prokic, J., & Huang, C.-R. (2023). ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection. In *Proceedings of the 4th international workshop on computational approaches to historical language change.* Singapore: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.lchange-1.10/

Choppa, T., Roth, M., & Schlechtweg, D. (2025). Predicting median, disagreement and noise label in ordinal word-in-context data. In *Proceedings of the 1st workshop on context and meaning–navigating disagreements in nlp annotations.* Abu Dhabi, UAE.

Chu, P. D. H. (2025). FuocChu_VIP123 at CoMeDi Shared Task: Disagreement ranking with xlm-roberta sentence embeddings and deep neural regression. In *Proceedings of the 1st workshop on context and meaning–navigating disagreements in nlp annotations.* Abu Dhabi, UAE.

Hätty, A., Schlechtweg, D., & Schulte im Walde, S. (2019). SURel: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the 8th Joint Conference on Lexical and Computational Semantics* (pp. 1–8). Minneapolis, MN, USA. Retrieved from https://www.aclweb.org/anthology/S19-1001/

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. SAGE Publications.

Kuklin, M., & Arefyev, N. (2025). Deep-change at CoMeDi: the cross-entropy loss is not all you need. In *Proceedings of the 1st workshop on context and meaning–navigating disagreements in nlp annotations.* Abu Dhabi, UAE.

Kurtyigit, S., Park, M., Schlechtweg, D., Kuhn, J., & Schulte im Walde, S. (2021, aug). Lexical Semantic Change Discovery. 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).* Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.acl-long.543/

References II

Kutuzov, A., & Pivovarova, L. (2021). Rushifteval: a shared task on semantic shift detection for russian. *Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference*.

Kutuzov, A., Touileb, S., Mæhlum, P., Enstad, T., & Wittemann, A. (2022, June). NorDiaChange: Diachronic semantic change dataset for Norwegian. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 2563–2572). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.lrec-1.274

Le, T. D., & Van, T. D. (2025). MMLabUIT at CoMeDi Shared Task: Text embedding techniques versus generation-based nli for median judgment classification. In *Proceedings of the 1st workshop on context and meaning–navigating disagreements in nlp annotations.* Abu Dhabi, UAE.

Leonardelli, E., Abercrombie, G., Almanea, D., Basile, V., Fornaciari, T., Plank, B., ... Poesio, M. (2023, July). SemEval-2023 task 11: Learning with disagreements (LeWiDi). In A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, & E. Sartori (Eds.), *Proceedings of the 17th international workshop on semantic evaluation (semeval-2023)* (pp. 2304–2318). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.semeval-1.314 doi: 10.18653/v1/2023.semeval-1.314

Liu, Z., Hu, Z., & Liu, Y. (2025). JuniperLiu at CoMeDi Shared Task: Models as annotators in lexical semantics disagreements. In *Proceedings of the 1st workshop on context and meaning–navigating disagreements in nlp annotations.* Abu Dhabi, UAE.

Loke, Y. X., Schlechtweg, D., & Zhao, W. (2025). ABDN-NLP at CoMeDi Shared Task: Predicting the aggregated human judgment via weighted few-shot prompting. In *Proceedings of the 1st workshop on context and meaning–navigating disagreements in nlp annotations.* Abu Dhabi, UAE.

Pilehvar, M. T., & Camacho-Collados, J. (2019, June). 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)* (pp. 1267–1273). Minneapolis, Minnesota: Association for Computational Linguistics. doi: 10.18653/v1/N19-1128

Rodina, J., & Kutuzov, A. (2020, dec). RuSemShift: a dataset of historical lexical semantic change in Russian. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics* (pp. 1037–1047). Barcelona, Spain (Online): International Committee on Computational Linguistics. Retrieved from https://aclanthology.org/2020.coling-main.90 doi: 10.18653/v1/2020.coling-main.90

Sakai, T. (2021). Evaluating evaluation measures for ordinal classification and ordinal quantification. 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)* (pp. 2759–2769).

References III

Sarumi, O. O., Welch, C., Seifert, C., Flek, L., & Schlötterer, J. (2025). Funzac at CoMeDi Shared Task: Modeling annotator disagreement from word-in-context perspectives. In *Proceedings of the 1st workshop on context and meaning–navigating disagreements in nlp annotations.* Abu Dhabi, UAE.

Schlechtweg, D. (2023). *Human and computational measurement of lexical semantic change* (Doctoral dissertation, University of Stuttgart, Stuttgart, Germany). Retrieved from http://dx.doi.org/10.18419/opus-12833

Schlechtweg, D., Cassotti, P., Noble, B., Alfter, D., Schulte Im Walde, S., & Tahmasebi, N. (2024, nov). More DWUGs: Extending and evaluating word usage graph datasets in multiple languages. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 14379–14393). Miami, Florida, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.emnlp-main.796

Schlechtweg, D., Choppa, T., Zhao, W., & Roth, M. (2025). The CoMeDi shared task: Median judgment classification & mean disagreement ranking with ordinal word-in-context judgments. In *Proceedings of the 1st workshop on context and meaning–navigating disagreements in nlp annotations.* Abu Dhabi, UAE.

Schlechtweg, D., Schulte im Walde, S., & Eckmann, S. (2018). 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* (pp. 169–174). New Orleans, Louisiana. Retrieved from https://www.aclweb.org/anthology/N18-2027/

Schlechtweg, D., Tahmasebi, N., Hengchen, S., Dubossarsky, H., & McGillivray, B. (2021, nov). 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* (pp. 7079–7091). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.567

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 88–103.

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., & Poesio, M. (2022, jan). Learning from disagreement: A survey. *J. Artif. Int. Res.*, *72*, 1385—1470. Retrieved from https://doi.org/10.1613/jair.1.12752 doi: 10.1613/jair.1.12752

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., . . . Bowman, S. R. (2019). Superglue: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd international conference on neural information processing systems.* Red Hook, NY, USA: Curran Associates Inc.

Wang, Z., Yu, A. W., Firat, O., & Cao, Y. (2021). *Towards zero-label language learning.* Retrieved from https://arxiv.org/abs/2109.09193

References IV

Zamora-Reina, F. D., Bravo-Marquez, F., & Schlechtweg, D. (2022). LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In *Proceedings of the 3rd international workshop on computational approaches to historical language change.* Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.lchange-1.16/