

A German WSC dataset comparing coreference resolution by humans and machines

Wiebke Petersen

Institute of Linguistics
Heinrich-Heine Universität
Düsseldorf, Germany
wiebke.petersen@hhu.de

Katharina Spalek

Institute of Linguistics
Heinrich-Heine Universität
Düsseldorf, Germany
katharina.spalek@hhu.de

Abstract

We present a novel German Winograd-style dataset for direct comparison of human and model behavior in coreference resolution. Ten participants per item provided accuracy, confidence ratings, and response times. Unlike classic WSC tasks, humans select among three pronouns rather than between two potential antecedents, increasing task difficulty. While majority vote accuracy is high, individual responses reveal that not all items are trivial and that variability is obscured by aggregation. Pre-trained language models evaluated without fine-tuning show clear performance gaps, yet their accuracy and confidence scores correlate notably with human data, mirroring certain patterns of human uncertainty and error. Dataset-specific limitations, including pragmatic reinterpretations and imbalanced pronoun distributions, highlight the importance of high-quality, balanced resources for advancing computational and cognitive models of coreference resolution.

1 Introduction

Coreference resolution is a central task in NLP (for a review see [Zhang et al., 2021](#)), with most work focusing on fine-tuning models for benchmark performance (e.g., [Wang et al., 2019](#)). In contrast, we directly compare the behavior of humans and pretrained language models (PTLMs) on a task requiring coreference resolution. Prior work shows that PTLMs encode coreference-relevant biases – such as preference for form similarity, recency, and grammatical agreement – when probed via contextual embeddings ([Sorodoc et al., 2020](#)), mirroring patterns found in human anaphora resolution (e.g., [Ariel, 2001](#); [Stevenson et al., 1995](#)). Yet for direct human-machine comparison, analyzing PTLM behavior during sentence processing offers more insight than diagnostic probing. Following [Ettinger](#)

(2020), we therefore assess PTLMs in a psycholinguistic setup.

We investigate how humans and PTLMs process German Winograd Schemas coreference problems designed to test commonsense reasoning and named after an example in [Winograd \(1972\)](#). The Winograd Schema Challenge (WSC) ([Levesque et al., 2012](#)) was proposed as a more demanding alternative to the Turing Test ([Turing, 1950](#)).¹ WSC items involve ambiguous pronouns whose resolution requires commonsense reasoning, and they are generally regarded as easy for humans but difficult for machines. A classic Winograd Schema (WS) consists of a pair of sentences differing only in a single critical word, that flips the intended referent of the pronoun. The classic task was to identify the correct antecedent:

- (1) Jane gave **Joan** candy because **she** was hungry.
Jane gave Joan candy because **she** wasn't hungry.
Who [was/wasn't] hungry? [() Jane; () Joan]

In parallel experiments, we compare how humans and machines differ in processing coreference. Specifically, we investigate (1) whether the same items are perceived as difficult by both groups, (2) which group performs better overall, and (3) whether model-based confidence measures (e.g., softmax probabilities) align with human self-assessed confidence ratings or response times. To ensure comparability, both groups perform *the same task on the same data*. Since we are interested in the linguistic knowledge encoded by pretrained models rather than in their capacity for fine-tuning, we deliberately refrain from additional training. Instead, we construct a dedicated dataset that allows

¹The original Turing Test (judging whether a conversational partner is human or not) has been criticized as too easy to pass through shallow mimicry rather than genuine understanding ([Weizenbaum, 1966](#)).

direct human–machine comparison on items reflecting tasks already encountered during pretraining.

2 Experiments

The key idea of our approach is to directly compare human and machine behavior on coreference resolution using a cloze-style task in German based on WSC items (see Fig. 1 for an example). The original WSC dataset (WSC273)² comprises 273 manually constructed WSC pairs like those in (1), where the task is to choose between two potential antecedents. The pairs are designed to meet three criteria: (a) the correct referent is unambiguous for humans; (b) resolution cannot rely solely on selectional restrictions; and (c) frequency-based heuristics are insufficient. Due to their difficulty and significance for machine translation and anaphora resolution, several larger WSC-style datasets have since been created. Among them, WinoGrande (Sakaguchi et al., 2021) is the most prominent, containing around 44,000 sentence pairs developed and validated via crowdsourcing. These are presented in cloze format, with the ambiguous pronoun replaced by a blank to be filled in, and two candidate antecedents provided as answer options. Reported human accuracy on these datasets typically exceeds 90% or even 95% (Kocijan et al., 2023).

Our approach is related to Abdou et al. (2020), who tested the robustness of humans and PTLMs on perturbed WSC items in cloze format (e.g., voice or tense changes), comparing majority vote (humans) and softmax predictions (PTLMs). While they focused on accuracy and stability, we go further by comparing confidence levels. To ensure comparability between the human and machine experiments we (1) avoid task priming by using fillers (humans) and no fine-tuning (machines), and (2) present structurally identical items to both groups.

2.1 Materials and task

We curated a set of 50 German WSC pairs satisfying two conditions: (i) each sentence contains two singular noun phrases of different grammatical gender and a gap to be filled with a nominative singular pronoun; (ii) both sentences differ in one critical word that determines the correct referent. Twenty-five pairs were randomly drawn from the 1m_en_de subset of MT-Wino-X (Emelin and Sennrich, 2021, here: Wino-X), a multilingual extension of Wino-

Grande for machine translation. The remaining 25 were translated from WSC273 using DeepL and manually revised to ensure grammaticality, naturalness, and a nominative singular pronoun gap. In cases where the gender of the two candidate referents did not differ, we adapted the referents accordingly (see (2), based on (1); for more details on the data adaption process see Appendix A).

- (2) Jan_{mas} gab Anne_{fem} Süßigkeiten, weil ____ *satt/hungrig* war.

The final dataset comprises 100 Winograd items (50 pairs): 42 with ‘sie’ (she) as the gold answer, 39 with ‘er’ (he), and 19 with ‘es’ (it). It was used in both experiments (humans and machines).

2.2 Human behavioral experiment


Using the dataset described in 2.1, we created ten experimental lists. Each list contained ten different WSC items (five from Wino-X, five from translated WSC273) and fifteen filler items, each presented as a cloze task (see Fig. 1). For no WSC pair both items belong to the same list. Filler items were designed to obscure the logical structure of the WSC-problems. As fillers we used sentences with only one potential antecedent, including entities with fixed (grammatical) gender (e.g., ‘der Tisch’) and ambiguous gender (e.g., brand names like ‘Nutella’, proper names like ‘Alex’ or foreign words like ‘Laptop’).

Participants selected the fitting German pronoun (*er*, *sie*, *es*) for each gap and rated their confidence on a 1–5 scale (see Fig. 1). Reaction times were recorded for both decisions. We tested 100 native German speakers (aged 18–55), collecting ten responses per WSC item. The experiment was conducted online using PsychoPy (Peirce, 2007) and distributed with via Clickworker. The experiment took 10-15 minutes, and participants received a small monetary compensation of 2.50€.

2.3 Pretrained language models (PTLMs) behavioral experiment

Our goal is to compare human and machine behavior on WSC items as directly as possible. We therefore evaluate PTLMs on the same cloze-style tasks used in the human experiment, without task-specific fine-tuning. This allows us to assess their inherent capabilities for coreference resolution based solely on their masked language modeling (MLM) pretraining.

²<https://www.tensorflow.org/datasets/catalog/wsc273>

Choose the fitting pronoun


Jan gab Anne Süßigkeiten, weil ____ hungrig war.

er

sie

es

How confident are you in your decision?

weak

•

•

•

•

strong

Figure 1: Human behavioral experiment: Pronoun choice and confidence rating, presented at two consecutive screens.

We include three BERT-based models: bert-base-german-cased, gbert-large (Chan et al., 2020), and xlm-roberta-large (Conneau et al., 2020). Each WSC item is converted into fill-mask format using the appropriate mask token. Softmax-normalized scores over the token vocabulary are interpreted as the model’s confidence in a token being the correct filler.

A key challenge is the mismatch between the tasks: humans are forced to choose one of three given pronouns (*er*, *sie*, *es*), while PTLMs predict freely from the entire vocabulary. To address this, we implement three configurations:

In the **pron**-configuration, only the three target pronouns are considered. The highest-scoring token among *er*, *sie*, and *es* defines the model’s prediction and its confidence. The score for the gold answer serves as the target confidence. This setup, however, disregards other high-scoring tokens that may function as pronoun synonyms in context.

The **topk**-configurations approximate the human task by including pronoun variants. The model’s top- k predictions are mapped to gendered pronoun classes (*masc*, *fem*, *neut*, *other*) using curated lists.³ The softmax scores of all top- k tokens belonging to each class are summed; the class with the highest total defines the model’s prediction and its confidence. We test $k = 10$ and $k = 1$, using either the summed gold-class score (**top10**) or the top-scoring gold-class token (**top1**) as target confidence.

Each configuration yields: (i) the model’s prediction, (ii) its correctness, (iii) its confidence in its given answer, and (iv) its target confidence (i.e., how strongly it favors the gold answer).

³E.g., *masc*: *der*, *er*, *dieser*, *jener*, etc.

3 Results and Discussion

In the behavioral experiment, we find a moderate inter-annotator agreement among humans ($\kappa = 0.562$), with only 28 items answered unanimously. This relatively low agreement is itself an important finding. First, it challenges the common assumption that WSC items are straightforward for humans and thus constitute a reliable benchmark for evaluating machines. Second, it raises concerns about the widespread practice of defining the human “gold” response via majority vote from as few as three annotators per item (see Kocijan et al., 2023, for a survey). The observed lack of high inter-annotator agreement suggests that majority votes based on larger samples may yield substantially different outcomes. Notably, for 21 items all three pronouns were chosen by at least one participant. At the same time, high agreement (≥ 7 of 10 participants selecting the same pronoun) was reached for 82 items, showing that while some items elicited highly consistent responses, a substantial number provoked genuinely ambiguous interpretations.

Table 1 summarizes model and human performance on our referential pronoun resolution task. Among models, GBERT-large and XLM-RoBERTa-large perform comparably (accuracy ≈ 0.56), both outperforming the smaller bert-base-german-cased (accuracy ≈ 0.53). Between configurations, accuracy remains largely stable, with top10 showing the highest target confidence, closely followed by pron, while top1 exhibits a notable drop. XLM-RoBERTa-large achieves slightly higher target confidence than GBERT-large and is therefore used in subsequent analyses. Overall, model accuracy is somewhat lower than previously reported results on English WSC data ($\sim 60\%$, Kocijan et al., 2023), likely due to the increased complexity of our task: models perform *free-form generation* over the full token vocabulary and humans choose among *three* options (rather than two in classic WSC).

Human performance is considerably higher than model performance, but also reveals striking variability. While majority vote accuracy is relatively high (0.87), individual accuracy is markedly lower (mean = 0.729). This challenges the assumption that WSC-style problems are trivial for humans and highlight the limitations of majority-based metrics, potentially masking individual uncertainty.

A breakdown by dataset reveals a strong quality gap: performance on Winograd-style expert-

Model	Accuracy			Target Conf.		
	top1	top10	pron	top1	top10	pron
XLM-RoBERTa	0.56	0.57	0.55	0.411	0.495	0.414
GBERT-large	0.56	0.56	0.56	0.397	0.481	0.410
BERT-base-german	0.53	0.52	0.53	0.342	0.412	0.350
Human (indiv.)	0.729			–		
Human (majority)	0.870			–		

Table 1: Model and human performance on Winograd cloze tasks. Accuracy refers to the proportion of correct predictions. Target confidence corresponds to the softmax score assigned to the gold token.

curated items (WSC273) is substantially higher than on Wino-X items, which are based on crowd-generated and machine translated data. Human majority vote accuracy is perfect on WSC273 but drops to 0.74 on Wino-X. Individual accuracy follows the same pattern (0.85 vs. 0.61). Model performance mirrors this trend (pron: 0.60 vs. 0.50), underscoring the importance of data curation.

In our analysis of human behavioral correlations, items with lower mean accuracy elicited longer mean response times ($r = -0.194$, $p < .001$) and lower mean confidence ratings ($r = 0.256$, $p < .001$). For the most extreme deciles mean accuracy raises from 0.65 for the slowest 10% of responses to 0.71 for the fastest, and from 0.49 for the lowest-rated items to 0.79 for the highest-rated ones, reinforcing the validity of these behavioral metrics. At the participant level, response time and confidence are themselves negatively correlated ($r = -0.142$, $p < .001$), indicating that individuals tended to take longer when less certain.

Comparing human and model behavior, we first note that all models predict the human majority vote more accurately than the gold answer (Table 5 in the appendix vs. Table 1). This suggests that models partially mirror human error patterns and produce judgments that align with aggregated human preferences.

Correlations between model confidence (measured as softmax scores for both the gold and given answer) and human behavioral measures are shown in Table 2. Unsurprisingly, due to the higher accuracy of human answers, model confidence in the gold answer correlates more strongly with human accuracy than confidence in the given answer. The same pattern is observed for correlations with human confidence ratings and response times, although the difference between gold and given answer is much smaller in this case. While our orig-

Model conf.	Acc.	Rating	RT
top1 (gold)	0.418	0.263	-0.245
top10 (gold)	0.410	0.280	-0.247
pron (gold)	0.486	0.307	-0.310
top1 (given)	0.222	0.260	-0.214
top10 (given)	0.120	0.225	-0.127
pron (given)	0.287	0.360	-0.253

Table 2: Pearson correlations between model confidence scores in gold and given answer and human measures (Acc. = correlation with mean human accuracy, Rating = correlation with mean human confidence ratings, RT = correlation with mean human reaction times). All p -values < 0.001 .

Model config.	V (indiv.)	V (maj. vote)
top1	0.433	0.503
top10	0.428	0.520
pron	0.441	0.503

Table 3: Cramér’s V between model predictions and human responses (individual and majority vote).

inal aim was to approximate model confidence via the given answer, gold-answer confidence ultimately shows the closest alignment with human behavior. Finally, despite its weak accuracy (see Table 1), the pron configuration shows the strongest correlations with human data across all configurations. Thus, despite lower correctness, its confidence estimates align more closely with human uncertainty and difficulty.

Complementing this, Cramér’s V analysis (Cramér, 1946) reveals moderate alignment between model outputs and individual human responses (Table 3). Again, pron shows the highest similarity to individual human response patterns, underscoring its ability to capture human-like behavior despite lower correctness. Additionally, model confidence in given answer correlates positively with human agreement (for pron: Spearman $r = 0.359$, $p < .001$; see Fig. 2 in the appendix), indicating that models are more confident when humans agree.

We observe several notable human and model error patterns. First, humans frequently select *es* to refer to an entire situation rather than one of the intended antecedents. Excluding items with wrong majority vote *es*, majority vote accuracy rises to 0.97 and individual accuracy to 0.80, indicating that many apparent ‘errors’ are due to pragmatic reinterpretation (see Appendix B for examples).

Second, models show markedly less variability

across paired items. While humans gave identical answers to both items of a pair in 10 cases, models did so over 35 times (e.g., 37 in pron). This suggests a tendency to being biased and ignore subtle contextual shifts that differentiate minimal pairs.

Third, both humans and models exhibit systematic biases in antecedent selection.⁴ Human responses show a slight preference for the *first antecedent* (515 vs. 424), while models exhibit a stronger bias toward the *second antecedent* (e.g., pron: 57 vs. 41), reflecting the recency bias observed in probing studies (Sorodoc et al., 2020). Model accuracy is higher when the correct referent is the first antecedent (e.g., pron: 0.59 vs. 0.54), while humans perform better when the correct answer is in the second position (0.74 vs. 0.81). This asymmetry is challenging to interpret, as pronoun types are not evenly distributed across positions and gender biases may influence performance. For instance, models perform best on *es* (pron: 0.68), followed by *sie* (0.57) and *er* (0.46), while humans show a minor difference between *er* (0.72) and *sie* (0.71), and a pronounced advantage for *es* (0.80).

4 Conclusion

We presented a novel German Winograd-style dataset and collected fine-grained human data, including accuracy, confidence ratings, and response times, with 10 participants per item.⁵ This resource provides a rich empirical basis for studying referential resolution in German and evaluating model behavior. Thereby our task setup is more challenging than previous WSC formulations: humans must choose among three pronouns, and models face open-ended generation over the entire vocabulary. Despite this, our results show clear human-machine performance gaps, alongside intriguing similarities in uncertainty and error patterns.

At the same time, our analysis reveals limitations in the dataset itself: some ‘errors’ reflect pragmatic reinterpretations rather than misunderstanding. Moreover, pronoun distribution is uneven across antecedent positions, suggesting room for improvement in future dataset design. Taken together, our findings reinforce the critical importance of high-quality, carefully constructed data for both cognitive and computational modeling of reference resolution.

⁴Note that the WSC pairs are balanced such that each antecedent position is correct equally often.

⁵The dataset is available from the authors upon request.

References

- Mostafa Abdou, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott, and Anders Søgaard. 2020. [The sensitivity of language models and humans to Winograd schema perturbations](#). In *Proc. of the 58th Annual Meeting of the ACL*, pages 7590–7604. ACL.
- Mira Ariel. 2001. [Accessibility theory: An overview](#). *Text Representation: Linguistic and Psycholinguistic Aspects*, 8(8):29–88.
- Branden Chan, Stefan Schweter, and Timo Möller. 2020. [German’s next language model](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Harald Cramér. 1946. *Mathematical Methods of Statistics*, volume 9 of *Princeton Mathematical Series*. Princeton University Press.
- Denis Emelin and Rico Sennrich. 2021. [Wino-X: Multilingual Winograd schemas for commonsense reasoning and coreference resolution](#). In *Proc. of the 2021 EMNLP*, pages 8517–8532, Punta Cana, Dominican Republic. ACL.
- Allyson Ettinger. 2020. [What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models](#). *TACL*, 8:34–48.
- Vid Kocijan, Ernest Davis, Thomas Lukasiewicz, Gary Marcus, and Leora Morgenstern. 2023. [The defeat of the Winograd schema challenge](#). *Artificial Intelligence*, 325:103971.
- Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. [The Winograd Schema Challenge](#). In *Proc. of the 13th International Conference on Principles of Knowledge Representation and Reasoning, KR’12*, pages 552–561. AAAI Press.
- Jonathan W. Peirce. 2007. PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1-2):8–13.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial Winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.
- Ionut-Teodor Sorodoc, Kristina Gulordava, and Gemma Boleda. 2020. [Probing for Referential Information](#)

in *Language Models*. In *Proc. of the 58th Annual Meeting of the ACL*. ACL.

Rosemary J Stevenson, Alexander WR Nelson, and Keith Stenning. 1995. *The role of parallelism in strategies of pronoun comprehension*. *Language and Speech*, 38(4):393–418.

A. M. Turing. 1950. *Computing machinery and intelligence*. *Mind*, LIX(236):433–460.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *Superglue: A stickier benchmark for general-purpose language understanding systems*. In *Neural Information Processing Systems*, pages 3266–3280.

Joseph Weizenbaum. 1966. ELIZA – a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

Hongming Zhang, Xinran Zhao, and Yangqiu Song. 2021. *A brief survey and comparative study of recent development of pronoun coreference resolution in English*. In *Proc. of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 1–11, Punta Cana, Dominican Republic. ACL.

A Details: Adaption of German WSC items for experiments

Half of the WSC items used in our experiments were drawn from the Wino-X dataset; the other half are adaptations of items from the original WSC273 set.

The `1m_en_de` subset of Wino-X is a subset of WinoGrande containing the English pronoun ‘it’. These were automatically translated into German, with ‘it’ replaced by a gap. Sentence pairs in which both versions required a pronoun of the same grammatical gender in German were excluded. For our study, we randomly selected 25 sentence pairs from this subset, ensuring only that the blank required a nominative singular pronoun. No further manual filtering or quality control was applied.

For the remaining 25 items, we randomly selected examples from WSC273 and translated them into German using DeepL. We then replaced the pronoun position with a blank and manually adjusted the sentences to (i) ensure grammatical fluency, (ii) require a nominative singular pronoun, and (iii) introduce two potential antecedents with different grammatical genders.

An example adapted from an original WSC pair is shown below:

- (3) a. original: The firemen arrived before the police because they were coming from so far away.
b. German adaption: Der Krankenwagen_{masc} kam vor der Polizei_{fem}, weil ____ so einen weiten Weg hatte.
The ambulance came before the police because ____ had such a long way.
- (4) a. original: The firemen arrived after the police because they were coming from so far away.
b. German adaption: Der Krankenwagen_{masc} kam nach der Polizei_{fem}, weil ____ so einen weiten Weg hatte.
The ambulance came before the police because ____ had such a long way.

The original English pair used the plural pronoun ‘they’, which was incompatible with our singular-pronoun setup. The automatic DeepL translation rendered the feminine singular nouns ‘fire department’ (‘Feuerwehr’) and ‘police’ (‘Polizei’). To introduce a gender contrast, we replaced ‘Feuerwehr’ with ‘Krankenwagen’ (‘ambulance’, masculine), enabling unambiguous pronoun resolution.

All item adaptations followed a similar procedure. Plural noun phrases were converted to singular, and gender-specific alternatives were introduced where necessary. Original WSC273 pairs typically involved ambiguous pronouns and names matched for gender. To ensure disambiguation via grammatical gender in German, we replaced personal names with frequent German names stereotypically associated with different genders.

B Human Majority Vote Errors

We begin by examining those 13 WSC items (out of 100) where the human majority vote diverged from the expected response. These instances highlight potential flaws in the item design, calling into question the claim that WSC-style problems are straightforward for humans. Fig. 2 shows that wrong majority votes occur across all agreement levels.

Several sources of confusion were identified:

Perspective shift: Some items allow for both pronouns to result in a coherent sentence by shifting

the perspective on the critical word.

- (5) “Die Frau kaufte eine Muschel_{fem}, um sie ins Aquarium_{neu} zu stellen, weil ____ schlicht aussah.”
majority vote: *sie* expected response: *es*
The woman bought a shell to put into the aquarium because ____ looked plain.

Both interpretations are plausible: either the shell is plain (*sie*) – and the women likes plain and simple things – or the aquarium is perceived as looking too plain without it (*es*).

Situational reference: Frequently, participants chose *es* to refer not to a noun, but to the entire situation. 50 times a participant answered *es* although neither the first nor the second antecedent had neuter gender.

- (6) Clara beschloss, Gemüse im Ofen_{masc} anstatt in der Mikrowelle_{fem} zu kochen, weil ____ das Gemüse saftiger schmecken ließ.
majority vote: *es* expected response: *er*
Clara decided to cook vegetables in the oven rather than the microwave because ____ made them taste juicier.

Here, *es* refers to the preparation process rather than a specific instrument.

This is the only example where both items in a WSC pair diverged from the expected response.

Gender error: German speakers are often uncertain about the grammatical gender of loanwords or less familiar nouns.

- (7) 3 Autos konnten in der Garage parken, aber nur 2 im Carport, da ____ kleiner war.
majority vote: *es* expected response: *er*
3 cars could park in the garage, but only 2 in the carport, because ____ was smaller.

Note that *Carport* is masculine, though even Wiktionary once mistakenly listed it as neuter.⁶

Complexity: Items can be complex due to too many potential antecedents.

- (8) Er konnte das Lenkrad in seinem Auto nicht vom Sitz aus erreichen, weil ____ zu niedrig war.
majority vote *es* expected response *er*
He couldn't reach the steering wheel in his car from his seat because ____ was too low.

⁶<https://de.wiktionary.org/wiki/Diskussion:Carport>

Item	E/M	Error
Die Frau kaufte eine Muschel, um sie ins Aquarium zu stellen, weil ____ schlicht aussah.	es/sie	persp.
Clara beschloss, Gemüse im Ofen anstatt in der Mikrowelle zu kochen, weil ____ das Gemüse knuspriger schmecken ließ.	er/es	sit.
Clara beschloss, Gemüse im Ofen anstatt in der Mikrowelle zu kochen, weil ____ das Gemüse saftiger schmecken ließ.	sie/es	sit.
Es war eine Herausforderung, den Kochtopf im Spülbecken zu waschen, da ____ flach war.	es/er	?
James ging in der Kälte mit einer Jacke anstelle eines Mantels zum Vorstellungsgespräch, weil ____ professionell aussah.	sie/es	sit.
Der Autor wollte den Monolog in der Geschichte verwenden, aber ____ war zu kurz.	sie/er	persp.
Ihre Beziehung verschlechterte sich auf dem Land, frischte jedoch in der Stadt auf, da ____ für sie eine so belebende Atmosphäre war.	sie/es	sit.
Ron wollte das Hühnerfleisch mit einer Gabel anstelle eines Messers zerkleinern, weil ____ besser funktionieren würde.	sie/es	sit.
Sie ging zum Strand und schwamm im Wasser, weil es so ein sonniger Tag war und ____ heiß war.	er/es	sit.
Eva stellte fest, dass die Pflanzen im Gewächshaus durch den Frost gediehen, während die im Garten starben, weil ____ kälter war.	er/es	sit.
Ich fühlte mich wohler, als ich meinen Freund im Haus küsste als im Park, weil ____ ein öffentlicher Ort war.	er/es	sit.
Er konnte das Lenkrad in seinem Auto nicht vom Sitz aus erreichen, weil ____ zu niedrig war.	er/es	compl.
3 Autos konnten in der Garage parken, aber nur 2 im Carport, da ____ kleiner war.	er/es	gender

Table 4: Items with diverging majority vote and expected response (E/M), including error classification (sit.: situational reference, persp.: perspective shift, gender: gender error, compl.: complexity, ?: unclear error source).

The item contains not just two, but four possible antecedents, namely *he*, *steering wheel*, *car*, and *seat*, for two of them it is plausible to be too ‘low’ in the context (he and seat), and only three are possible by the selectional restrictions of ‘niedrig’ (low), namely car, seat and steering wheel.

Table 4 summarizes all 13 cases. Notably, all problematic items come from the Wino-X dataset, not our adapted WSC273 items. This may be due to the fact that WSC273 problems were carefully crafted and reviewed by experts, while Wino-X items stem from crowdsourced WinoGrande problems and may lack this level of precision.

C Additional Tables and Graphs

Model	top1	top10	pron
XLNet-RoBERTa	0.640	0.650	0.640
GBERT-large	0.620	0.650	0.640
BERT-base-german-cased	0.580	0.590	0.600

Table 5: Accuracy of each model configuration in predicting the human majority vote.

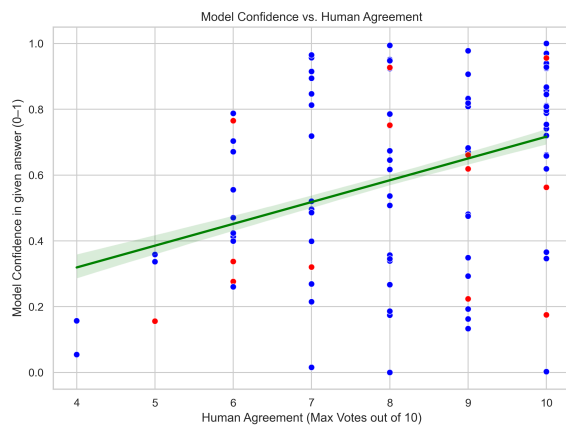


Figure 2: Model confidence (pron configuration) as a function of human agreement (maximum number of votes for a pronoun out of 10). Items where the majority vote is incorrect are shown in red.