Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)

This isn't the bias you're looking for: Implicit causality, names and gender in German language models

Sina Zarrieß and **Hannes Gröner** and **Torgrim Solstad** and **Oliver Bott**
Bielefeld University
Linguistics Department
{sina.zarriess,hannes.groener,torgrim.solstad,oliver.bott}@uni-bielefeld.de

Abstract

To assess whether neural language models capture discourse-level linguistic knowledge, previous work has tested whether they exhibit the well-known implicit causality (IC) bias found in various interpersonal verbs in different languages. Stimuli for analyzing IC in computational and psycholinguistic experiments typically exhibit verb arguments with different genders. In this paper, we revisit IC in German neural language models, analyzing gender and naming bias as a potential source of confusion. Indeed, our results suggest that IC biases in two existing models for German are weak, unstable, and behave in unexpected and unsystematic ways, when varying names or gender of verb arguments.

1 Introduction

In recent years, large-scale pretrained neural language models (PLMs) have not only become an important component in modeling many NLP tasks (Devlin et al., 2019; Liu et al., 2019; Sanh et al., 2019; Lewis et al., 2020; Brown et al., 2020), but the models themselves have turned more and more into the subject of linguistic analysis and probing: One prominent line of work has investigated undesired social biases, e.g. gender or racial biases, that PLMs inherit from the large and often unmoderated resources for training (Bordia and Bowman, 2019; Blodgett et al., 2020; Meade et al., 2022). Another line of work has examined the linguistic knowledge and desirable biases captured in PLMs, ranging from morphological, syntactic and semantic to discourse-related probing tasks (Belinkov and Glass, 2019; Ettinger, 2020).

In this work, we built upon a series of recent papers that investigated a desirable linguistic bias in PLMs: the implicit causality bias (Upadhye et al., 2020; Davis and van Schijndel, 2020; Kementchedjhieva et al., 2021). Implicit Causality (IC) is a property of a wide range of interpersonal verbs like *annoy*, which display a preference for establishing coreference to one of the verb's argument over the other in explanations:

(1) Peter annoyed Mary because ... ...

When asked to continue a sentence like (1), human subjects have a strong preference towards referring to *Peter*, as in *because he sang loudly*, attributing the implicit cause to the stimulus argument (the subject of *annoy*, in this case). In order to be able to experimentally assess such next-mention biases, studies in (computational) psycholinguistics commonly use stimuli where the verb's arguments mismatch in their gender, so that continuations with a female or male pronoun unambiguously refer to the subject or object of the main clause.

Previous studies on testing IC in PLMs designed stimuli with two NPs in different genders, generating language model prompts with varying names and orders, carefully balanced for gender (Upadhye et al., 2020; Kementchedjhieva et al., 2021). However, they did not explicitly examine the potential interactions with underlying gender bias in PLMs, despite the fact that this a well-known and widely discussed phenomenon in recent work in NLP.

In this paper, we revisit the IC bias for two German language models, BERT and GPT-2, based on Solstad and Bott (2022)'s experimental data. We analyze PLMs' predicted continuations of prompts with an interpersonal verb and two gender-mismatched arguments followed by a connective, as shown in example (1). As in previous studies, we vary and balance prompts for the names and gender of verb arguments and introduce a further condition that manipulates the form of names: next to first names like *Anna, Paul*, we test surnames like *Herr Müller (Mr. Müller), Frau Fischer (Ms. Fischer)*, which in German carry accusative case

marking (*Herrn Müller*). Our analysis shows that the manipulation of names' form and gender uncovers various inconsistencies in the continuations predicted by German PLMs for IC prompts.

2 Background

2.1 IC: Implicit Causality and Consequentiality

As discussed by Solstad and Bott (2022), psychological verbs like the stimulus-experiencer (SE) verb *annoy* and the experiencer-stimulus (ES) verb *fear* display biases for establishing coreference to one of the verbs arguments in the context of explanation and consequence. In explanation contexts (introduced by the connective *because*), continuations have a strong referential bias to re-mention the stimulus argument. In consequence contexts (introduced by the connective *and so*), however, an equally strong re-mention bias towards the mention of the experiencer argument is observed. As shown in Examples (2)-(3), this leads to a mirror subject bias pattern: the ES-verb in Example (2) has a bias towards the subject in explanation and towards the object in consequence contexts (the preferred continuation is shown in brackets), whereas the SE-verb in Example (3) shows the complementary bias pattern:

(2) a. Mary fears Peter because... [he] is always so aggressive.
b. Mary fears Peter and so ... [she] tries to avoid him.

(3) a. Mary annoys Peter because... [she] is so ignorant.
b. Mary annoys Peter and so... [he] acted rather impolite.

In psycholinguistic sentence completion studies, participants generally receive a prompt including the connective. In their continuations they typically provide reference to the biased argument (in square brackets).

In the following, we will subdivide IC into Implicit Causality (I-Caus) and Implicit Consequentiality (I-Cons). For I-Caus, Solstad and Bott (2022) found a subject-bias for SE verbs and an object-bias for ES verbs with 87.4% and 4.0% subject coreference in continuations, respectively. I-Cons continuations displayed the exact opposite biases with 4.8% subject continuations for SE and 77.9% subject continuations for ES verbs. The opposite I-Caus and I-Cons biases were reflected by an almost perfect negative correlation between I-Caus and I-Cons biases ($r = -0.94, p < .001$) making I-Caus and I-Cons biases of the two psych-verb classes a very interesting testing ground for language models.

Upadhye et al. (2020) used a similar set-up to ours, distinguishing between IC1 and IC2 verbs as well as explanations and consequences. These correspond to SE and ES verbs as well as the I-Caus and I-Cons condition in our setting. Kementchedjhieva et al. (2021) investigate IC in PLMs, but do not discuss mirror biases in their set-up. In general, these previous studies obtained mixed but overall rather promising results in favour of predictions congruent with human-like next-mention biases. Upadhye et al. (2020) find that two English PLMs (Transformer-XL, GPT-2) are not sensitive to manipulations of connectives in IC contexts, but that GPT-2 assigns higher probability to subject-referring pronouns when the respective interpersonal verb exhibits a strong subject bias in human completions, and vice versa for object-referring pronouns. Kementchedjhieva et al. (2021) test a wider range of English PLMs and find that bidirectional models in particular show a moderate to strong correlation with human completions in IC contexts. They also report results on German and Spanish, with German BERT achieving moderate correlations with human IC bias data.

2.2 Gender Bias and Implicit Causality

Bias studies often employ two different-gender names to ease the assessment of coreference with subject or object arguments, i.e. there is a subject bias when the prnoun is male and the first argument of the main verb is a male first name. Typically, the order of male and female referents is included as a counterbalancing factor (e.g., *Peter/Mary annoyed Mary/Peter*) to exclude that gender biases interfere with coreference biases. For instance, a gender bias would be observed if the subject bias for SE verbs in I-CAUS context is less strong when the stimulus is female as compared to male.

Mostly, as in Solstad and Bott's (2022) study, no gender effects have been found. However, Ferstl et al. (2011) did find the proportions of coreference for IC ('because') to be skewed towards male referents. Importantly, Ferstl et al. observed an interaction with participant gender to the extent that male participants were more likely to attribute the

cause to the male referent, irrespective of subject or object position. In light of the well-known and widely attested gender bias in neural language models and word embeddings (Blodgett et al., 2020), we argue that the lack of analysis of gender bias in the context of implicit causality constitutes an interesting research gap, that the current study is aiming to fill.

3 Experiments

3.1 Materials

We based our study on I-Caus and I-Cons in German on Experiment 1 in Solstad and Bott (2022). The experiment employed a 2×2(×2) within-participants and within-items design manipulating the factors VERB CLASS (German stimulus-experiencer vs. experiencer-stimulus verbs) and CONNECTIVE (*weil* 'because' vs. *sodass* 'and so'). They chose these two connectives because of their optimal syntactic parallelism. Differently from *daher* or *deswegen* ('therefore') the chosen connectives both select for subordinate sentences with pronouns typically immediately following the connective (similar to the English examples in (2)/(3)). This is a very important prerequisite for probing pronoun production. The form *sodass* is nowadays the most frequent variant of this connective (as suggested by the google books ngram viewer), while forms such as *so dass*, *sodaß* and *so daß* are more infrequent in use.

In addition, GENDER ORDER (male>female vs. female>male) was included as a counterbalancing factor. Solstad and Bott (2022) included 20 stimulus-experiencer and 20 experiencer-stimulus verbs, which were chosen for their stable and pronounced biases. Items were constructed according to a *name*$_1$ *verb-ed name*$_2$ *connective* scheme in line with the above design. Verbs were paired in items matching them semantically as closely as possible. The resulting 20 items in eight conditions were distributed to four list using a Latin Square design, with proper names chosen from publicly available lists of the most frequent first names in Germany.[1] Sentence completions were elicited from 52 participants (39 female; 13 male).

3.2 Language Model Prompts

We use Solstad and Bott (2022)'s experimental items to generate German prompts to be completed by the language models. As in the above examples, prompts consist of a simple sentence introducing the verb, the verb's arguments and the connective:

(4) a. Peter langweilte Marie, sodass ...
Peter bored Mary and so ...

b. Frau Müller sorgte sich um Herrn Schmidt, weil ...
Mrs. Müller was worried about Mr. Schmidt because ...

In contrast to the English Examples (2)-(3), the German Example (4-a) allows for both subject-before-object (SVO) as well as object-before-subject (OVS) interpretations, i.e. *Peter* could be the stimulus or experiencer of the event. The ambiguity does not arise when the arguments are realized as surnames, as in Example (4-b), due to the accusative marking on the word *Herr*. Solstad and Bott (2022) explicitly annotated whether their human participants had assigned an OVS interpretation to the prompts and observed that against this potential concern overwhelmingly SVO interpretations were chosen in more than 95% of the cases. In our study, we assume that the first argument always refers to the subject. In future work, it may be of interest to estimate the amount of OVS interpretations assigned by PLMs, too.

We balanced the prompts according to the following properties:

ES vs. SE Our set of verbs divides into 20 experiencer-stimuli verbs (ES, see Example (4-b)) and 20 stimuli-experiencer verbs (SE, see Example (4-a)).

I-CAUS vs. I-CONS For each verb, we created templates with the connective *weil* 'because' for implicit causality (I-Caus, see Example (4-b)) and *sodass* 'and so' for implicit consequentiality (I-Cons, see Example (4-a)).

First names vs. surnames For each template, we created prompts using five surnames and five first names, e.g., *Herr Schmidt, Paul, Anna*. In each case, both verb arguments were instantiated witht the same type of name.

[np1] We balanced the prompt set for each verb such that the gender of the first argument (i.e. the subject) in the sentence is male/female in 50% of the cases. In Example (4-a), [np1] is male (m), in Example (4-b) it is female (f).

Taken together, we obtain a set of 100 prompts for each of the 40 verbs.

[1] Full materials at https://osf.io/5ewbd/

Bias type	NP-type	BERT	GPT-2
overall	all	0.581	0.560
	firstn	0.556	0.568
I-CAUS	all	0.576	0.548
	firstn	0.503	0.577
I-CONS	all	0.585	0.571
	firstn	0.609	0.559

Table 1: Completion sensitivity for BERT and GPT-2 in I-CAUS and I-CONS contexts, with all types of names and first names (firstn) only

3.3 Models and Metrics

We used two German language models to generate continuations of the set of prompts: (i) the pre-trained DBMZ German **GPT-2** model[2], and (ii) the cased DBMZ German **BERT** model [3], a fully bidirectional model.

From these models, we obtain the likelihood assigned to the continuations *er* (*he*) and *sie* (*she*). We calculate the subject bias for human and model continuations and use the metrics of Prediction Accuracy and Completion Sensitivity from Ettinger (2020).

Completion Sensitivity For each prompt, there is a presumed bias on either the first or the second noun phrase. A pronoun is said to be congruent with the bias if it refers to the noun phrase specified by the bias. Completion Sensitivity scores are calculated as the percentage of prompts where the predicted pronoun is congruent with the bias.

Prediction Accuracy (Acc@2) Prediction Accuracy scores are calculated as the percentage of prompts, where *he* or *she* are among the top 2 continuations.

Subject Bias Subject bias scores are calculated as the percentage of prompts where the pronoun referring to the subject ([np1]) has a higher probability than the pronoun referring to the object ([np2]).

[2] https://huggingface.co/dbmdz/german-gpt2

[3] https://huggingface.co/dbmdz/bert-base-german-cased

4 Results

Table 1 shows completion sensitivity results aggregated for all types of verbs and names. To ease comparison with previous studies, we also report aggregated results on prompts with first names only. In general, these scores suggest that both language models have a weak but seemingly consistent tendency to generate continuations congruent with human biases, i.e. more than 50% of the predictions are congruent in I-Caus and I-Cons conditions. However, results shown in Table 2 suggest that generated continuations are much less consistent than scores in Table 1 may lead us to expect.

As shown in the more detailed breakdown in Table 2, continuations predicted by GPT-2 generally exhibit a strong object bias (low subject bias scores in all conditions), a finding that aligns well with Kementchedjhieva et al. (2021)'s results on German PLMs. This object bias is less strong, however, in some conditions where the subject is female, but only when it is additionally realized as a first name (I-Caus/SE and I-Cons/SE+ES). Moreover, we note that GPT-2 prediction accuracy (Acc@2) drops substantially for all I-Cons/SE verbs, as well as for some I-Caus/ES verbs with female subjects or surname subjects. For the I-Cons/ES condition with female surname subjects, the prediction accuracy is close to 0. This indicates that GPT-2 does not only fail in capturing next-mention biases for interpersonal verbs in our data, but rather fails to compute reliable representations of complex entity names and clauses embedded with *sodass* (*and so*).

Continuations predicted by BERT do not exhibit any systematic object or subject bias across conditions, nor do they exhibit biases that align well with human continuations. For instance, in I-Caus contexts with ES verbs, BERTs predictions display an object bias (in line with humans), except when the subject is female and realized as a surname. In I-Caus contexts with SE verbs, BERTs predictions display an object bias for first name (not in line with humans), but a subject bias for surnames (which would be in line with humans). Similar patterns arise in I-Cons contexts: for ES verbs, predictions tend towards an object bias, except when the subject is a female surname (94% subject bias). Additionally, prediction accuracies in I-Cons contexts drop systematically and dramatically across different verb and name types. Again, this indicates that the model fails to compute reliable representations of prompts ending in *sodass* (*and so*), which,

Bias type	V-type	NP-type	[np1]	BERT Acc@2	BERT Subject Bias	GPT-2 Acc@2	GPT-2 Subject Bias	Human Subject Bias
I-CAUS	ES	firstn	m	0.814	0.118	0.926	0.004	0.06
			f	0.922	0.148	0.826	0.092	0.02
		surn	m	0.898	0.264	0.872	0.000	
			f	0.954	0.520	0.462	0.000	
	SE	firstn	m	1.000	0.200	1.000	0.008	0.885
			f	1.000	0.080	1.000	0.396	0.862
		surn	m	0.998	0.564	1.000	0.074	
			f	0.928	0.818	1.000	0.002	
I-CONS	ES	firstn	m	0.578	0.398	1.000	0.134	0.81
			f	0.568	0.436	0.992	0.368	0.748
		surn	m	0.528	0.462	0.958	0.330	
			f	0.156	0.944	1.000	0.004	
	SE	firstn	m	0.522	0.344	0.736	0.000	0.05
			f	0.336	0.054	0.820	0.266	0.045
		surn	m	0.698	0.518	0.778	0.000	
			f	0.744	0.640	0.072	0.000	

Table 2: Top-2 prediction Accuracy (Acc@2), and Subject Bias for BERT and GPT-2 predictions, and human continuations for different contexts (I-Caus/I-Cons, Experiencer-Stimuli (ES) Stimuli-Experiencer (SE) verbs, NPs with first names (firstn) and surnames (surn). Human scores for prompts using surnames are not available.)

in German, is less frequent than *weil* (*because*).

Discussion Generally, our results indicate that the large-scale German PLMs we tested in this study are not able to compute reliable discourse-level representations of our prompts that are abstract enough to capture next mention bias for interpersonal verbs, regardless of the realization of the names in verbs' arguments. This mirrors Abdou et al. (2020)'s findings on Winograd schema perturbations, showing that language models are sensitive to minimal changes in prompts that do not affect human understanding. Our results also support proposals to improve the modeling of names and entities in neural language models (Ji et al., 2017; Févry et al., 2020; Holgate and Erk, 2021). Concerning gender bias, BERT's continuations show tendencies towards a female bias when NPs are realized as surnames, which may be related to the fact that German *sie* is ambiguous and can refer to female singular and plural entities.

5 Conclusion

We have investigated implicit causality and consequentiality biases in two German PLMs. We find that GPT-2 shows a strong object bias, which is weaker for prompts where the verb arguments are realized as surnames and the subject's gender is female. BERT does not exhibit any systematic next-mention bias for I-Caus and I-Cons conditions when gender and name type are varied. Thus, none of the models show evidence for human-like next-mention biases in explanation or consequence contexts. In line with Abdou et al. (2020), we conclude that perturbation and variation of experimental stimuli is an important tool when testing PLMs on data collected in psycholinguistic studies with humans.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7590–7604, Online. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Forrest Davis and Marten van Schijndel. 2020. Discourse structure interacts with reference but not syntax in neural language models. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 396–407, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Evelyn C. Ferstl, Alan Garnham, and Christina Manouilidou. 2011. Implicit causality bias in English: a corpus of 300 verbs. *Behavior Research Methods*, 43(1):124–135.

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4937–4951, Online. Association for Computational Linguistics.

Eric Holgate and Katrin Erk. 2021. "politeness, you simpleton!" retorted [MASK]: Masked prediction of literary characters. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 202–211, Groningen, The Netherlands (online). Association for Computational Linguistics.

Yangfeng Ji, Chenhao Tan, Sebastian Martschat, Yejin Choi, and Noah A. Smith. 2017. Dynamic entity representations in neural language models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1830–1839, Copenhagen, Denmark. Association for Computational Linguistics.

Yova Kementchedjhieva, Mark Anderson, and Anders Søgaard. 2021. John praised Mary because _he_? implicit causality bias and its interaction with explicit cues in LMs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4859–4871, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Torgrim Solstad and Oliver Bott. 2022. On the nature of implicit causality and consequentiality: the case of psychological verbs. *Language, Cognition and Neuroscience*, 37:1–30. Ahead-of-print version.

Shiva Upadhye, Leon Bergen, and Andrew Kehler. 2020. Predicting reference: What do language models learn about discourse models? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 977–982, Online. Association for Computational Linguistics.