Proceedings of the Society for Computation in Linguistics

Volume 6 Article 8

6-1-2023

Does a Neural Model Understand the De Re / De Dicto Distinction?

Gaurav Kamath
McGill University, gaurav.kamath@mail.mcgill.ca

Laurestine Bradford
McGill University, laurestine.bradford@mail.mcgill.ca

Does a neural model understand the *de re / de dicto* distinction?

Gaurav Kamath[1,*]
gaurav.kamath
@mail.mcgill.ca

Laurestine Bradford [1,2,*]
laurestine.bradford
@mail.mcgill.ca

[1] McGill University and Mila
[2] Centre for Research on Brain, Language and Music

Abstract

Neural network language models (NNLMs) are often casually said to "understand" language, but what linguistic structures do they really learn? We pose this question in the context of *de re / de dicto ambiguities*. Nouns and determiner phrases in intensional contexts, such as belief, desire, and modality, are subject to referential ambiguities. The phrase "Lilo believes an alien is on the loose," for example, has two interpretations: one (*de re*) in which she believes a specific entity which happens to be an alien is on the loose, and another (*de dicto*) in which she believes some unspecified alien is on the loose. In this paper we confront an NNLM with contexts producing *de re / de dicto* ambiguities. We use coreference resolution to investigate which interpretive possibilities the model captures. We find that while RoBERTa is sensitive to the fact that intensional predicates and indefinite determiners each change coreference possibilities, it does not grasp how the two interact with each other, and hence misses a deeper level of semantic structure. This inquiry is novel in its cross-disciplinary approach to philosophy, semantics and NLP, bringing formal semantic insight to an active research area testing the nature of NNLMs' linguistic "understanding."

1 Introduction

Modern neural net language models (NNLMs) are often publicized as "understanding" language, which can belie a lack of knowledge about the nature of the linguistic structures they truly capture (Bender and Koller, 2020). Consequently, there has been much interest in probing NNLMs' sensitivity to theoretical linguistic structures, an area which Baroni (2021) calls *linguistically-oriented deep net analysis* (LODNA). Such analysis often uses psycholinguistic methods to give NNLMs acceptability tasks similar to those one would give to a human (Warstadt et al., 2019). Existing work has primarily measured NNLMs' ability to capture syntactic structures (Bacon, 2020; Linzen and Baroni, 2021; Warstadt et al., 2019), though a few semantic phenomena, such as the causative-inchoative alternation, have also been investigated (Warstadt et al., 2019).

Fine-grained semantic distinctions present unique difficulties for LODNA. It can be challenging to pose the right problems to test NNLM knowledge of subtle meaning distinctions; for example, see (Tsiolis, 2020)'s discussion in the context of quantifier scope ambiguity. Nonetheless, fine-grained semantic distinctions are crucial to modern theories of semantic structure, and it is therefore important to find out how well NNLMs "understand" them. One such subtle meaning difference lies in the *de re* and *de dicto* interpretations of noun phrases in intensional contexts.

The *de re / de dicto* distinction, made notable by Quine (1956) among others, refers to two distinct kinds of interpretations of noun phrases that arise from intensional contexts in natural language. Such contexts include belief, desire, and modality. The statement "Lilo believes an alien is on the loose," for example, has two interpretations. Under one interpretation (*de re*), Lilo believes a specific entity that just so happens to be an alien (say, Stitch) is on the loose. Lilo herself (as is the case in *Lilo and Stitch* (Sanders and DeBlois, 2002)) need not know that Stitch is an alien for the statement to be true. Under the other interpretation (*de dicto*) Lilo believes that some unspecified alien, whatever it may be, is on the loose. Unlike the *de re* interpretation, no alien needs to actually exist for the statement to be true under this interpretation.

De re / de dicto ambiguities have traditionally been treated in the philosophy and semantics literature as scope ambiguities, where each interpretation arises out of a modal or intensional operator outscoping, or being outscoped by, another

[*] Equal contribution.

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 69-84.
Amherst, Massachusetts, June 15-17, 2023

quantifier (see (Keshet and Schwarz, 2019) for an overview). For example:

De re: $\exists x[\text{alien}_{w_0}(x) \land \forall w'$ $[\text{BEL}_{w_0}(\text{Lilo},w') \Rightarrow \text{on-the-loose}_{w'}(x)]]$

De dicto: $\forall w'[\text{BEL}_{w_0}(\text{Lilo},w') \Rightarrow$ $\exists x[\text{alien}_{w'}(x) \land \text{on-the-loose}_{w'}(x)]]$[1]

NNLMs, however, lack any similar formal system of representation, since all meaning representation is contained within numerical embeddings and weights. This provides further theoretical motivation to investigate whether NNLMs are capable of discerning *de re* / *de dicto* ambiguities, and whether they show any bias towards either interpretation. If NNLMs are capable of making these distinctions, it would suggest not only that they are capable of mimicking human-like fine-grained semantic distinctions, but also that numerical vectors are rich enough to capture deep formal structure. We thus believe that the capacity of NNLMs to discern *de re* / *de dicto* ambiguities has strong implications for both semantics and NLP.

Therefore, we investigate whether current powerful language models can interpret NPs in intensional contexts in both *de re* and *de dicto* senses. We will do so by framing the problem as one of coreference resolution.

2 Related Work

As NNLMs have become increasingly successful at a range of natural language tasks in recent years, there has been much discussion of the capacity of such models to "understand" language. While this use of the term is misleading (Bender and Koller, 2020), it has spurred research into the ability of NNLMs to pick up on theoretical, often complex linguistic structures.

Most of this LODNA work has focused on syntactic structures. For overviews of such work, see (Baroni, 2021; Bender and Koller, 2020; Linzen and Baroni, 2021). The present paper differs from this body of work, however, in that we address a semantic, rather than a syntactic, phenomenon.

Although not as much, there has also been work in LODNA on semantics. For example, some progress has been made in measuring the degree to which NNLMs encode compositionality (Ettinger et al., 2018; Shwartz and Dagan, 2019; Jawahar et al., 2019; Yu and Ettinger, 2020, 2021; Bogin et al., 2022) and systematicity (Lake and Baroni, 2018; Goodwin et al., 2020; Kim and Linzen, 2020). Researchers have also studied the capacity of NNLMs to capture more specific, fine-grained semantic phenomena, including monotonicity (Yanaka et al., 2019), the causative-inchoative alternation (Warstadt et al., 2019), negation (Ettinger et al., 2018; Ettinger, 2020; Kim et al., 2019; Richardson et al., 2020), and quantification (Kim et al., 2019; Richardson et al., 2020).

Natural language understanding (NLU) benchmarks also have the opportunity to test models' grasp of theoretical semantic structures. Most large collections of NLU benchmarks focus on performance of specific tasks (such as sentiment analysis and question answering) rather than abstract linguistic knowledge (Liang et al., 2020; Ruder et al., 2021; Dumitrescu et al., 2021; Ham et al., 2020; Khashabi et al., 2020; Park et al., 2021; Rybak et al., 2020; Seelawi et al., 2021; Wilie et al., 2020; Yao et al., 2021). Indeed, Bowman and Dahl (2021) have argued that targeting specific linguistic knowledge can hinder performance of NNLMs on NLP tasks.

Nevertheless, some NLU benchmarks overlap with LODNA in addressing certain theoretical semantic structures. In particular, the benchmarks discussed in (Xia and Van Durme, 2021) all assess models' semantically-informed coreference resolution capability, as do the collection of benchmarks following the Winograd Schema (Levesque et al., 2012; Kocijan et al., 2020), which includes some large benchmark sets like those mentioned above (Wang et al., 2019a,b; Xu et al., 2020; Shavrina et al., 2020). A benchmark nearer to the spirit of LODNA is proposed in (Yanaka et al., 2021). This paper directly relates generation of NNLM test cases to theoretical semantic structures. The authors use such structures to create tests for NNLMs' compositional generalization of logical operators, modifiers, and embedded clauses. Finally, in the class of NLU benchmarks, the work of (Ribeiro et al., 2021) is nearest to our own investigation. Here, the author proposes templates that can be filled in to create probes of NNLMs' capability with a variety of structures. These structures include antonymy, temporal ordering, negation, and coreference. Note that none of the previous work

[1] While other equivalent formulations of the logical forms of such sentences are present in the literature, we choose to adopt the same notation as (Zhang and Davidson, 2021), on account of its conciseness and simplicity.

assesses modality or intensionality. In the present work, we employ a template-like scheme for generating test cases that assess NNLMs' behaviour in intensional contexts.

We focus on the *de re* / *de dicto* distinction. Since being highlighted in recent times by (Quine, 1956), *de re* / *de dicto* ambiguities have been the subject of extensive work in philosophy and semantics. For an overview, see (Keshet and Schwarz, 2019). Most of this work focuses on of how to formally represent intensional contexts (Fodor, 1970; Tichý, 1971; Montague, 1973; Lewis, 1979; Von Fintel and Heim, 2011); specific points of focus include scope (Keshet, 2008, 2010), (Elliott, 2022), modality (Plantinga, 1969; Fine, 1978), and even tense (Ogihara, 1996; Kauf and Zeijlstra, 2018). For all this work on the theory of *de re* / *de dicto* ambiguities, however, there is a dearth of experimental work on the distinction. The work reported in (Zhang and Davidson, 2021) therefore stands out for its quantitative experimental approach. The authors conduct an study directly measuring whether English speakers demonstrate any preference towards *de re* or *de dicto* readings. Their results suggest that speakers accept *de dicto* interpretations more robustly than *de re* interpretations.

To our knowledge, there has been no similar attempt to situate *de re* / *de dicto* ambiguities in the context of NNLMs. Williamson et al. (2021) present an amendment to Abstract Meaning Representation (AMR), a graphical meaning representation language, which allows it to encode *de re* / *de dicto* ambiguities as scope ambiguities. This marks perhaps the closest recent work on these ambiguities in a NLP context. AMR, however, is an artificial meaning representational language, and therefore of a different type than the meaning representation of an NNLM. Our work directly looks for *de re* / *de dicto* ambiguities in NNLMs' behaviour.

3 Model

In all experiments, we use a version of the RoBERTa (Liu et al., 2019) masked language model already fine-tuned for the SuperGLUE Winograd Schema Challenge task (Levesque et al., 2012; Wang et al., 2019a). This is because: (i) our method of distinguishing *de re* from *de dicto* interpretations centers on recognizing coreference, which this model does well at, scoring 89% on the SuperGLUE WSC task (while for comparison, OpenAI's few-shot GPT-3 scores 80.1%) (Wang et al.); and (ii) this model proved most straightforward to access and work with. We directly access and work with this model using Meta AI's fairseq library (Ott et al., 2019).

4 Dataset and evaluation metric

4.1 Dataset

We generate a dataset of test sentences that consist of a matrix subject, an intensional verb with sentential complement, an embedded subject, and an embedded intransitive verb. The matrix subject is always *John* or *Mary*, and the embedded subject is always a noun phrase. All of the test cases have either the form in Figure 1a, as in the example *John believes that a dentist is singing*, or the form in Figure 1b, as in the example *John wants a dentist to be singing*. The choice between these structures simply depends on whether the matrix verb requires a finite or non-finite tense in its complement.

We simultaneously generate a dataset of sentences which are similar to the above, but with a perceptual verb instead of an intensional verb. These therefore have the form in Figure 1c, as in the example *John sees a dentist singing*. Note that perceptual verbs have been analyzed by a few in the literature as also being intensional (e.g. Bourget, 2017); for sentences with perceptual verbs, we therefore have the perceptual verbs take direct objects as their arguments (as in *John sees a dentist singing*), rather than clauses (as in *John sees that a dentist is singing*), so as to minimize the possibility of intensional interpretations of the perceptual verbs.

Sentence templates are generated from the schemata in Figure 1 with every possible combination of: *John* or *Mary* in the matrix subject, a verb from the list in Appendix A.3 in the matrix verb, a noun from the list in Appendix A.1 in the embedded subject, and a verb from the list in Appendix A.2 in the embedded verb.

In addition to manipulating whether the matrix verb is intensional, we manipulate the determiner of the embedded subject. We generate alternations between the indefinite determiner 'a'/'an', as in *Mary believes that a dentist is smiling*, and the deictic determiner 'that', as in *Mary believes that that dentist is smiling*. The indefinite 'a'/'an' should give rise to a *de re* / *de dicto* ambiguity. The deictic 'that' should, in theory, only allow for a *de re* interpretation, since it must refer to an entity already present in the world of discourse.

[MatrixSubject]	[MatrixVerb] *that*	[EmbeddedSubject] *is*	[EmbeddedVerb]
John	*believes*	*an editor*	*walking*
Mary	*accepts*	*a dentist*	*singing*
	deduces	*a baker*	*shouting*
	...	...	...

(a) Intensional sentences with finite-tensed complements.

[MatrixSubject]	[MatrixVerb]	[EmbeddedSubject] *to be*	[EmbeddedVerb]
John	*wants*	*an editor*	*walking*
Mary	*wishes for*	*a dentist*	*singing*
	requires	*a baker*	*shouting*
	...	...	...

(b) Intensional sentences with non-finite-tensed complements.

[MatrixSubject]	[MatrixVerb]	[EmbeddedSubject]	[EmbeddedVerb]
John	*sees*	*an editor*	*walking*
Mary	*observes*	*a dentist*	*singing*
	hears	*a baker*	*shouting*
	...	...	...

(c) Perceptual sentences.

Figure 1: Schemata for generating test data

We handpick 48 matrix verbs (36 intensional and 12 perceptual), randomly select 60 embedded nouns from a handpicked list of 204, and randomly select 30 embedded verbs from a handpicked list of 51[2]. The resultant dataset contains a total of 345,600 unique sentences with the configurations shown in Figure 1 (although the total size of dataset is larger, for reasons explained in the following section). 259,200 of these are sentences with intensional verbs, and the remaining 86,400 are sentences with perceptual verbs.

4.2 Evaluation

The availability of the embedded NP as an anaphoric antecedent depends on whether it is interpreted *de re* or *de dicto*. Consequently, for each generated sentence, we post-pend three different fixed sentences: (i) *I met [pronoun]*, (ii) *I greeted [pronoun]*, and (iii) *I liked [pronoun]*[3]. We then use a tweaked version of the WSC-finetuned RoBERTa model's in-built `disambiguate_pronoun` function to obtain the scores the model assigns at the *[pronoun]* position to each possible coreferent (i.e. the main subject or the embedded subject)[4].

Under the *de dicto* reading, the embedded NP should not be able to corefer with a subsequent phrase, as under this reading it is interpreted solely within the intensional context. By contrast, under the *de re* reading, the embedded NP should be able to corefer with a subsequent phrase, as under this reading it is interpreted outside the intensional context.

In intuitive terms, using the example *Mary believes that a lawyer is shouting*, under the *de dicto* interpretation, the lawyer is only specified in Mary's beliefs, rather than the speaker's world of reference. But the subsequent post-pended sentence is evaluated with respect to the speaker's world of reference, and not Mary's beliefs. So, the pronoun token in the post-pended sentence should not be able to refer to the embedded NP. Under a *de re* interpretation, however, the lawyer is specified in the speaker's world of reference. So it remains accessible for coreference in the post-pended sentence.

Therefore, we should be able to assess the perfor-

[2] We randomly select subsets of these lists, instead of using the entire handpicked lists, due to concerns of dataset size and excessive compute requirements with little obvious *a priori* benefit of using the complete lists.

[3] This triples the final size of our dataset, bringing it to 1,036,800.

[4] In this process, the model doesn't actually make use of the token in the position it predicts for. We therefore use the *[pronoun]* token as a placeholder for what is in effect a masked position, as using RoBERTa's actual <*mask*> token led to issues with the code.

mance of the masked language model at detecting the *de re* / *de dicto* ambiguity by comparing the scores it assigns to the matrix or the embedded subject at the pronoun position. For example, in *Mary believes that a dentist is singing. I met [pronoun]*, we compare the scores assigned to the possible coreferents *Mary* and *a dentist* at the pronoun position[5]. We use three separate post-pended sentences to try to ensure that the effects we see are not the result of any one specific verb in the follow-up sentence.

Scores assigned to the matrix subject should be higher for test sentences where the matrix verb is intensional and the embedded subject has an 'a'/'an' determiner. These are the contexts that give rise to the possible *de dicto* interpretation which would exclude the embedded subject from coreference. By contrast, the relative scores for the matrix and embedded subject should be closer to equal in cases that only admit a *de re* interpretation. This includes all cases with a 'that' determiner or where the matrix verb is perceptual (i.e. non-intensional).

5 Results and Discussion

5.1 Results

To quantify the model's coreference choice at the pronoun position, we study the difference between the score assigned to the matrix subject (e.g. *John*) and that assigned to the embedded subject (e.g. *an actor*); we call this difference *matrix subject bias*. Figure 2 shows the empirical effect of matrix verb type and determiner type on matrix subject bias. We see an overall increase in matrix subject bias in intensional contexts and in contexts where the embedded subject has the determiner 'a' or 'an'. The difference between intensional and perceptual contexts is slightly smaller when the embedded subject has determiner 'a' or 'an'.

In order to study the effects of interest while marginalizing over other manipulations and over random variability, we fit a linear mixed-effects model with formula below (random effects specified in brackets).

$$\begin{aligned}
&\text{Matrix Subject Bias} \sim \\
&1 + \text{Determiner} * \text{Matrix Verb Type} \\
&\qquad + \text{Followup Verb} + \text{Matrix Subject} \\
&+ (1 + \text{Determiner} + \text{Matrix Subject} \\
&\qquad + \text{Followup Verb} \mid \text{Matrix Verb}) \\
&+ (1 + \text{Determiner} * \text{Matrix Verb Type} \\
&\qquad + \text{Followup Verb} + \text{Matrix Subject} \\
&\qquad \mid \text{Embedded Verb}) \\
&+ (1 + \text{Determiner} * \text{Matrix Verb Type} \\
&\qquad + \text{Followup Verb} + \text{Matrix Subject} \\
&\qquad \mid \text{Embedded Subject})
\end{aligned}$$

[5]The implementation of coreference resolution in the model we use is such that a span such as *a dentist* is not penalized simply for being longer than a single token like *Mary*.

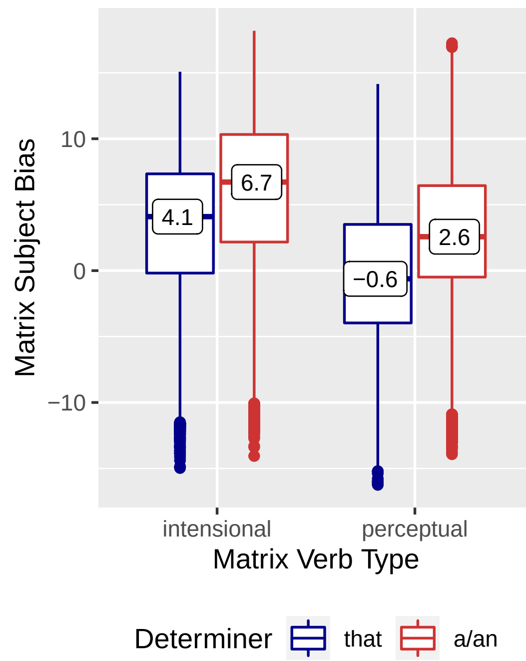


Figure 2: Boxplot with whiskers to 1.5IQR showing distribution of matrix subject bias by determiner and matrix verb type.

The full results are reported in Tables 1 and 2. The model confirms the overall trend in Figure 2. Averaged across all conditions, there is a bias towards matrix subjects of 3.27 points (df=71.91, t=6.96, p<0.001). Sentences with perceptual matrix verbs show 2.58 points lower matrix subject bias than those with intensional matrix verbs (df=78.16, t=-5.03, p<0.001), and sentences with with determiner 'a'/'an' show 2.89 points higher matrix subject bias than those with determiner 'that' (df=92.78, t=11.92, p<0.001). The effect of verb type is smaller in indefinite ('a'/'an') determiner contexts than deictic ('that') contexts by 0.52 points, but this is not statistically significant (df=72.83,t=1.44,p=0.152).

There is considerable variability in both effects according to embedded verb and embedded subject, and variability in the determiner effect according to matrix verb, embedded verb, and embedded subject

Coefficient	$\hat{\beta}$	$\mathbf{SE}(\hat{\beta})$	df	t	p
Intercept	3.27	0.47	71.91	6.96	< 0.001
Determiner = ‘a/an’	2.89	0.24	92.78	11.92	< 0.001
Matrix Verb Type = ‘perceptual’	-2.58	0.51	78.16	-5.03	< 0.001
Matrix Subject = ‘Mary’	-1.27	0.17	89.18	-7.65	< 0.001
Followup Verb = ‘liked’ (vs. ‘greeted’)	-0.25	0.26	102.91	-0.97	0.333
Followup Verb = ‘met’ (vs. 0.5(‘liked’+‘greeted’))	-1.12	0.13	96.10	-8.93	< 0.001
Interaction Determiner:Matrix Verb Type	0.52	0.36	72.83	1.44	0.152
Marginal $R^2 = 0.21$, Conditional $R^2 = 0.65$, $n = 1036800$,					
Groups: Matrix Verb (48); Embedded Verb (30); Embedded Subject (60)					

Table 1: A regression table showing fixed effects, goodness of fit, and test statistics for the linear mixed-effects model in Section 5.1. Degrees of freedom and p-values estimated using the Satterthwaite approximation. Predictor levels were coded as ± 0.5, except Followup Verb coded with Helmert contrasts.

Group	Term	Variance	SD
Matx. Verb	Intercept	1.13	1.49
	Determiner	0.89	0.94
	Matx. Subj	0.05	0.22
	Foll. Verb Cont.1	1.12	1.05
	Foll. Verb Cont.2	0.23	0.48
Emb. Verb	Intercept	3.92	1.98
	Determiner	0.76	0.87
	Matx. Verb Type	2.02	1.42
	Matx. Subj	0.17	0.42
	Foll. Verb Cont.1	0.84	0.92
	Foll. Verb Cont.2	0.22	0.47
	Det.:Matx. Type	0.80	0.90
Emb. Subj	Intercept	1.92	1.39
	Determiner	0.50	0.71
	Matx. Verb Type	0.79	0.89
	Matx. Subj	1.25	1.12
	Foll. Verb Cont.1	0.88	0.93
	Foll. Verb Cont.2	0.21	0.46
	Det.:Matx. Type	0.38	0.62
Residual		10.09	3.18

Table 2: A table showing fitted random effects of the model specified in Section 5.1, as well as residual variance.

(Table 2). Nonetheless, the overall trend is clear.

See Appendix B for an overview of additional trends which do not bear on the main research question.

5.2 Discussion

From these results, it is clear that both verb type (intensional or non-intensional) and determiner type (indefinite or deictic) have statistically significant effects on the relative scores the language model assigns to different possible anaphoric referents.

Intensional verbs yield higher matrix subject bias than non-intensional, perceptual verbs, when all other variables are held constant. This is in line with our predictions, as intensional verbs allow for *de dicto* readings that block the embedded subject from coreference.

In addition, indefinite determiners yield higher matrix subject bias than deictic determiners. This is also in line with our predictions, as indefinite determiners are more amenable to *de dicto* readings that block the embedded subject from coreference. However, the interaction between these two factors is not statistically significant. This goes against our predictions, as deictic determiners should bias the reader toward *de re* readings no matter what, so the matrix verb effect should diminish when the determiner is ‘that’.

These results are positive evidence that neural language models can be sensitive to the effect of intensional predicates on *de re* / *de dicto* ambiguities, and therefore to intensionality more broadly. However, the lack of interaction suggests that there is something deeper that RoBERTa misses. It captures the effects of verb intensionality and deictic determiners; however, it does not capture the correct result of combining the two. By contrast, a formal-theoretical model of intensional verbs’ and of determiners’ meanings would lead naturally to the correct inference that deictic determiners facilitate *de re* readings regardless of matrix verb.

Some other results are also worth mentioning, shown in more detail in Appendix B. As seen in Table 1 and Figure 4b, the matrix subject bias is very similar when the followup verb is *liked* or *greeted*, but lower in a statistically significant way when it is *met*. The reason for this effect is not known. Whether the matrix subject is *Mary* or *John* has a statistically significant effect on matrix subject bias; holding other variables constant, setting the matrix subject to *Mary* instead of *John* yields a lower matrix subject bias. Given the propensity for large language models to be gender-biased in various ways (Lu et al., 2020; Vig et al., 2020; Charlesworth et al., 2021), this is perhaps not surprising.

6 Conclusion

In this paper, we investigate the capacity of a neural language model, a version of RoBERTa fine-tuned for coreference resolution, to identify *de re* / *de dicto* ambiguities that arise in intensional contexts. We find evidence suggesting that such models are indeed sensitive to the ambiguity-generating effects of intensional predicates and the ambiguity-resolution effects of deictic determiners, but find no evidence that this sensitivity extends to the interaction between intensional predicates and embedded determiners.

Our approach is also subject to some limitations that invite further research. Our range of test data is tightly constrained in its syntactic and broad semantic structure. This is deliberate, as we hoped to isolate the semantic effects of intensional predicates and determiners from the confounding factors of syntactic form and broader semantic context. However, the downside of this approach is that our findings may not generalize across more varied forms of language. Similarly, our choice of perceptual verbs as the counterpart to intensional verbs was the result of their shared syntactic properties, which allowed for substitution while holding all other variables (including sentence structure) virtually unchanged. One possibility, however, is that the effects we find between intensional and perceptual verbs are dependent on the latter's being specifically perceptual verbs, and do not represent a difference between intensional and non-intensional verbs more generally. Finally, in this paper, we work with only one model. Other models with different architecture or pretraining may have produced different results.

Clearly, a broader study of the capacity of neural models to capture intensional effects such as *de re* / *de dicto* ambiguities requires a wider set of data and experimental setups. We hope that this inquiry spurs further research to that end.

7 Code

Code and data for this project are available at `https://github.com/laurestine/nnlm-de-re-de-dicto`.

8 Acknowledgements

The authors would like to thank Siva Reddy for his guidance, as well as Chris Potts and the anonymous reviewers for their feedback on earlier versions of this work. The CRBLM is funded by the Government of Quebec via the Fonds de Recherche Nature et Technologies and Société et Culture.

References

Geoffrey I. Bacon. 2020. *Evaluating linguistic knowledge in neural networks*. Ph.D. thesis, UC Berkeley.

Marco Baroni. 2021. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Ben Bogin, Shivanshu Gupta, and Jonathan Berant. 2022. Unobserved local structures make compositional generalization hard. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2731–2747, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Bourget. 2017. Intensional perceptual ascriptions. *Erkenntnis volume*, 82.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Tessa E. S. Charlesworth, Victor Yang, Thomas C. Mann, Benedek Kurdi, and Mahzarin R. Banaji. 2021. Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, 32(2):218–240. PMID: 33400629.

Stefan Daniel Dumitrescu, Petru Rebeja, Beata Lorincz, Mihaela Gaman, Andrei Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, Luciana Morogan, George Dima, Gabriel

Marchidan, Traian Rebedea, Madalina Chitez, Dani Yogatama, Sebastian Ruder, Radu Tudor Ionescu, Razvan Pascanu, and Viorica Patraucean. 2021. LiRo: Benchmark and leaderboard for Romanian language tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Patrick D Elliott. 2022. A flexible scope theory of intensionality. *Linguistics and Philosophy*, 46:333–378.

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.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *COLING*.

Kit Fine. 1978. Model theory for modal logic. part i – the de re/de dicto distinction. *Journal of Philosophical Logic*, 7(1):125–156.

Janet Dean Fodor. 1970. *The Linguistic Description of Opaque Contexts*. Ph.D. thesis, MIT.

Emily Goodwin, Koustuv Sinha, and Timothy J. O'Donnell. 2020. Probing linguistic systematicity. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1958–1969, Online. Association for Computational Linguistics.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. *CoRR*, abs/2004.03289.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Carina Kauf and Hedde Zeijlstra. 2018. Towards a new explanation of sequence of tense. In *Semantics and Linguistic Theory*, volume 28, pages 59–77.

Ezra Keshet. 2008. *Good intensions: paving two roads to a theory of the de re / de dicto distinction*. Ph.D. thesis, MIT.

Ezra Keshet. 2010. Split intensionality: A new scope theory of de re and de dicto. *Linguistics and Philosophy*, 33(4):251–283.

Ezra Keshet and Florian Schwarz. 2019. De re/de dicto. *The Oxford handbook of reference*, pages 167–202.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabadi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, Erfan Sadeqi Azer, Niloofar Safi Samghabadi, Mahsa Shafaei, Saber Sheybani, Ali Tazarv, and Yadollah Yaghoobzadeh. 2020. ParsiNLU: A suite of language understanding challenges for Persian. *CoRR*, abs/2012.06154.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Vid Kocijan, Thomas Lukasiewicz, Ernest Davis, Gary Marcus, and Leora Morgenstern. 2020. A review of Winograd schema challenge datasets and approaches. *CoRR*, abs/2004.13831.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *13th International Conference on the Principles of Knowledge Representation and Reasoning, KR 2012*, Proceedings of the International Conference on Knowledge Representation and Reasoning, pages 552–561. Institute of Electrical and Electronics Engineers Inc. 13th International Conference on the Principles of Knowledge Representation and Reasoning, KR 2012 ; Conference date: 10-06-2012 Through 14-06-2012.

David Lewis. 1979. Attitudes de dicto and de se. *Philosophical Review*, 88(4):513–543.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1):195–212.

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. *CoRR*, abs/1907.11692.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. *Gender Bias in Neural Natural Language Processing*, pages 189–202. Springer International Publishing, Cham.

Richard Montague. 1973. The proper treatment of quantification in ordinary english. In Patrick Suppes, Julius Moravcsik, and Jaakko Hintikka, editors, *Approaches to Natural Language*, pages 221–242. Dordrecht.

Toshiyuki Ogihara. 1996. *Tense, attitudes, and scope*, volume 58 of *Studies in Linguistics and Philosophy*. Springer Science & Business Media.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won-Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Tae Hwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Eunjeong Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. *CoRR*, abs/2105.09680.

Alvin Plantinga. 1969. De re et de dicto. *Noûs*, 3(3):235–258.

Willard Quine. 1956. Quantifiers and propositional attitudes. *Journal of Philosophy*, 53:177–187.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2021. Beyond accuracy: Behavioral testing of NLP models with checklist (extended abstract). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4824–4828. International Joint Conferences on Artificial Intelligence Organization. Sister Conferences Best Papers.

Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8713–8721.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. KLEJ: comprehensive benchmark for Polish language understanding. *CoRR*, abs/2005.00630.

Chris Sanders and Dean DeBlois. 2002. *Lilo & Stitch*. Walt Disney Pictures.

Haitham Seelawi, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. 2021. ALUE: Arabic language understanding evaluation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 173–184, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Tatiana Shavrina, Alena Fenogenova, Anton A. Emelyanov, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. *CoRR*, abs/2010.15925.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Pavel Tichý. 1971. An approach to intensional analysis. *Noûs*, 5(3):273–297.

Konstantinos Christopher Tsiolis. 2020. Quantifier scope disambiguation.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Kai Von Fintel and Irene Heim. 2011. Intensional semantics. *Unpublished Lecture Notes*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE leaderboard. Available at `https://super.gluebenchmark.com/leaderboard` (2022/04/18).

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. *SuperGLUE: A Stickier Benchmark for General-Purpose Language*

Understanding Systems. Curran Associates Inc., Red Hook, NY, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. *CoRR*, abs/2009.05387.

Gregor Williamson, Patrick Elliott, and Yuxin Ji. 2021. Intensionalizing Abstract Meaning Representations: Non-veridicality and scope. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 160–169, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patrick Xia and Benjamin Van Durme. 2021. Moving on from OntoNotes: Coreference resolution model transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5241–5256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liang Xu, Xuanwei Zhang, Lu Li, Hai Hu, Chenjie Cao, Weitang Liu, Junyi Li, Yudong Li, Kai Sun, Yechen Xu, Yiming Cui, Cong Yu, Qianqian Dong, Yin Tian, Dian Yu, Bo Shi, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. *CoRR*, abs/2004.05986.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, and Kentaro Inui. 2021. SyGNS: A systematic generalization testbed based on natural language semantics. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 103–119, Online. Association for Computational Linguistics.

Yuan Yao, Qingxiu Dong, Jian Guan, Boxi Cao, Zhengyan Zhang, Chaojun Xiao, Xiaozhi Wang, Fanchao Qi, Junwei Bao, Jinran Nie, Zheni Zeng, Yuxian Gu, Kun Zhou, Xuancheng Huang, Wenhao Li, Shuhuai Ren, Jinliang Lu, Chengqiang Xu, Huadong Wang, Guoyang Zeng, Zile Zhou, Jiajun Zhang, Juanzi Li, Minlie Huang, Rui Yan, Xiaodong He, Xiaojun Wan, Xin Zhao, Xu Sun, Yang Liu, Zhiyuan Liu, Xianpei Han, Erhong Yang, Zhifang Sui, and Maosong Sun. 2021. CUGE: A Chinese language understanding and generation evaluation benchmark. *CoRR*, abs/2112.13610.

Lang Yu and Allyson Ettinger. 2021. On the interplay between fine-tuning and composition in transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2279–2293, Online. Association for Computational Linguistics.

Lang-Chi Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *EMNLP*.

Yuhan Zhang and Kathryn Davidson. 2021. *De re* interpretation in belief reports: An experimental investigation. In *Experiments in Linguistic Meaning*, volume 1. Linguistic Society of America.

A Lexical items used in stimuli

A.1 Embedded Subjects

We used the following nouns as embedded subjects, sampled randomly from a list of English nouns denoting professions and types of person:

actor
administrator
ambassador
architect
assistant
baker
bartender
boy
chancellor
clerk
clown
controller
cook
cooper
count
courier
dancer
dealer
dentist
designer
dictator
diver
drummer

economist
editor
emperor
engineer
farmer
girl
governor
guard
guitarist
historian
journalist
king
lady
lawyer
lieutenant
lobbyist
lord
magician
manager
mayor
merchant
model
negotiator
novelist
painter
philosopher
producer
psychiatrist
publisher
queen
rabbi
solicitor
spy
supervisor
treasurer
waiter
woman

A.2 Embedded Verbs

We used the following embedded intransitive verbs, sampled randomly from a list of English intransitive verbs denoting activities.

arriving
coughing
cringing
crying
dying
hiccuping
kneeling
limping
lying
moving
panicking
partying
praying
resting
running
screaming
shouting
sighing
singing
sitting
smiling
smoking
sneezing
standing
sweating
swimming
talking
walking
waving
working

A.3 Matrix Verbs

We used the following intensional matrix verbs, meant to be as wide an array of intensional verbs as possible:

accepts
aims for
anticipates
assumes
believes
concludes
conjectures
deduces
demands for
desires for
doubts
dreads
expects
fears
feels
figures
gathers
guesses
hopes
imagines
intends for
knows
maintains
needs
presumes

reckons
requires
supposes
surmises
suspects
thinks
trusts
understands
wants
wishes for
worries

We used the following perceptual matrix verbs, meant to be as wide an array of perceptual verbs as possible:

catches sight of
detects
glimpses
hears
notices
observes
overhears
perceives
sees
spots
views
watches

B Data distribution details

This appendix contains additional details, not directly relevant to our research questions, about patterns in matrix and embedded subject scores.

Figure 3 shows the raw distribution of matrix and embedded subject scores. Matrix subject scores are generally higher than embedded subject scores.

Figures 4a and 4b show distribution of matrix subject bias for each matrix subject and for each followup. We see that 'met' yields considerably lower matrix subject bias than other followup verbs, while matrix subjects of John are preferred as coreferents more than matrix subjects of Mary.

Figure 5 shows distribution of matrix subject bias for each determiner-syntactic frame pair. We see that the two intensional-verb frames pattern together in the way indicated in the main text: they have higher matrix subject bias than the perceptual-verb frame, and all three frames show higher matrix subject bias with indefinite determiners.

We next computed the raw effect of determiner, the raw effect of intensional matrix verb, and their

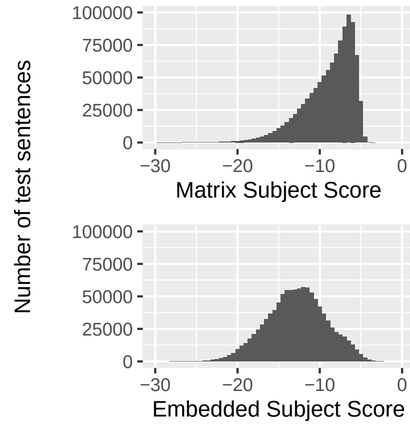
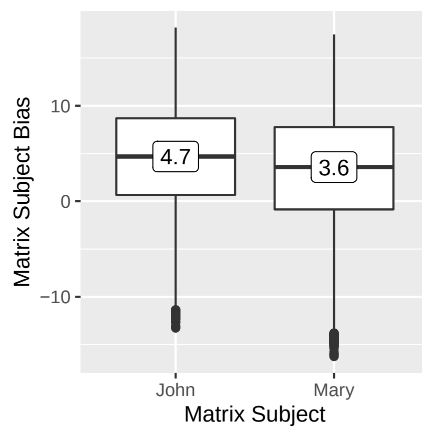
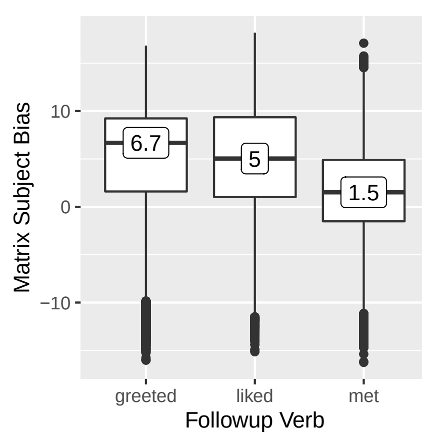


Figure 3: Histograms showing the raw distribution of matrix and embedded subject scores.



(a)



(b)

Figure 4: Boxplot with whiskers to 1.5IQR showing the distribution of matrix subject bias for each matrix subject and for each followup verb.

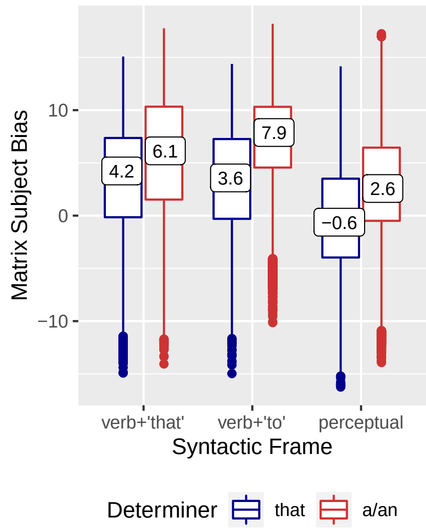


Figure 5: Boxplot with whiskers to 1.5IQR showing the distribution of matrix subject bias by syntactic frame and determiner.

interaction separately for each possible matrix subject, embedded subject, embedded verb, and followup verb. The results are shown in Figure 6. Raw effects are computed as differences of means, and the raw interaction is a difference of differences of means. We see that the overall positive effect of indefinite determiner and intensional matrix verb is a trend across the bulk of data points, and is not merely the result of a few outliers. The lack of interaction between these two effects is also consistent. Figure 7 shows the pattern that test sentence frames with "liked" as a followup verb have a higher effect of determiner than those with other followup verbs, but we see that the effect of an indefinite determiner on matrix subject bias is still positive in general.

Finally, Figures 8, 9, and 10 show variability in matrix subject score and embedded subject score depending on the specific choice of embedded subject (Figure 8), embedded verb (Figure 9), and matrix verb (Figure 10). This variability is quite high, with some lexical items in each case showing almost no matrix subject bias, and others showing quite a lot. Aside from our deliberate manipulation of intensionality, it is unclear what else drives this variability.

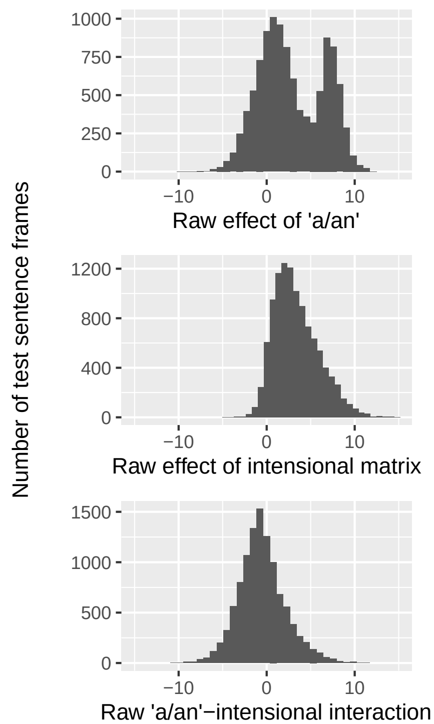


Figure 6: Sentence frames plotted by their raw effect of indefinite determiner (difference in matrix subject bias between instances of that frame with indefinite and deictic determiners), raw effect of intensional matrix verb (difference in mean matrix subject bias between instances of that frame with an intensional and perceptual matrix verb), and raw interaction of these two effects (difference-of-differences between the aforementioned subgroups).

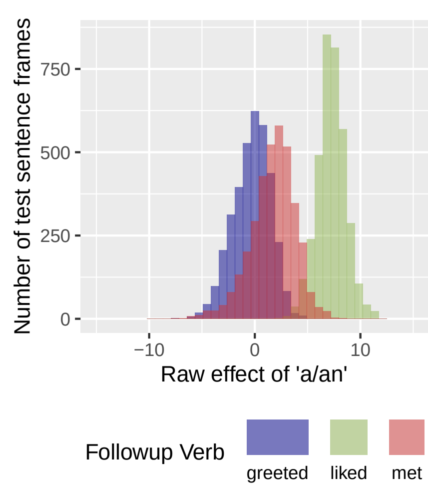


Figure 7: Sentence frames plotted by their raw effect of indefinite determiner, colored by followup verb.

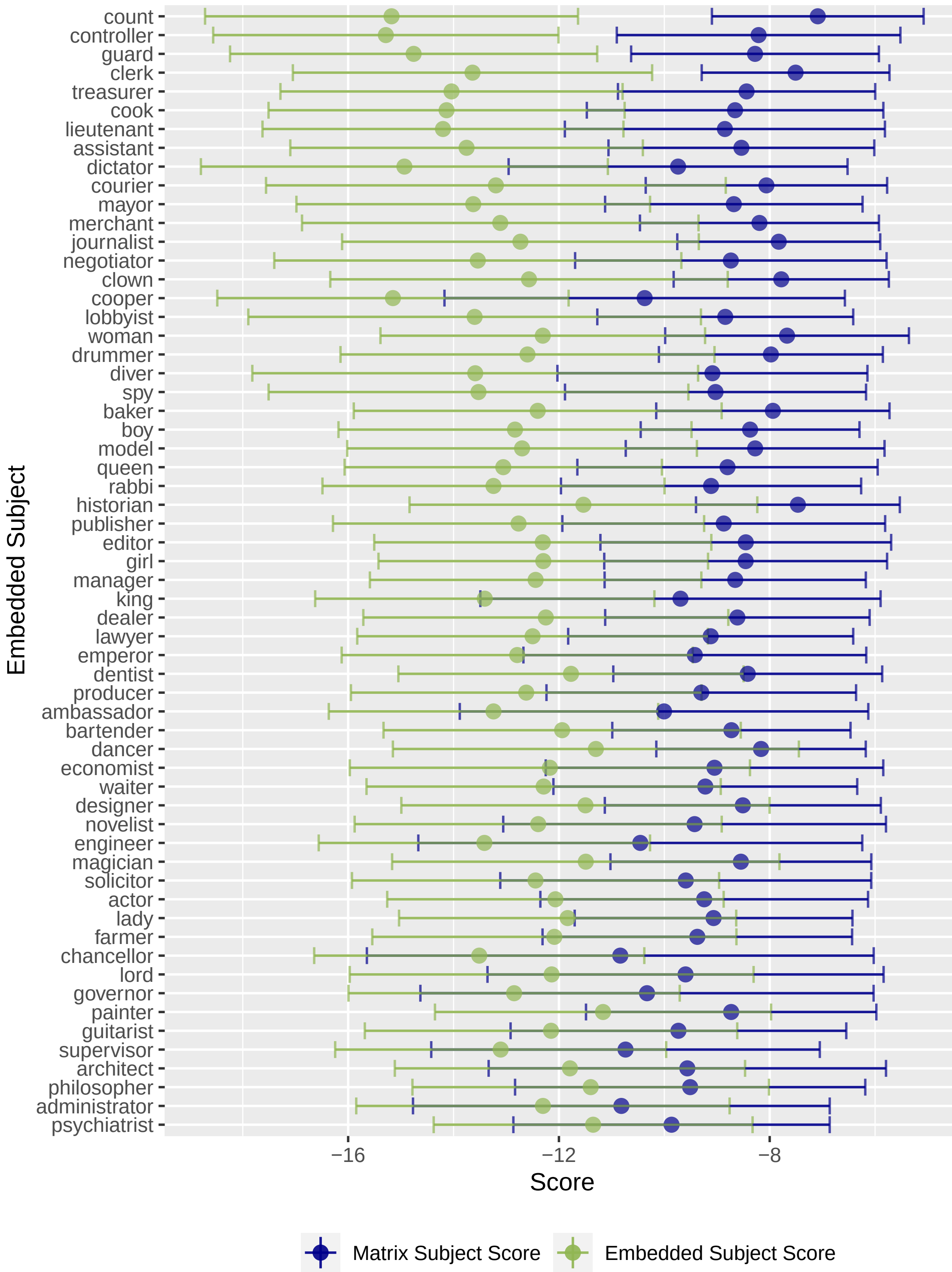


Figure 8: Error bar plot showing mean matrix subject score and embedded subject score for stimuli with each embedded subject. Rows are ordered by matrix subject bias. Error bars show standard deviation.

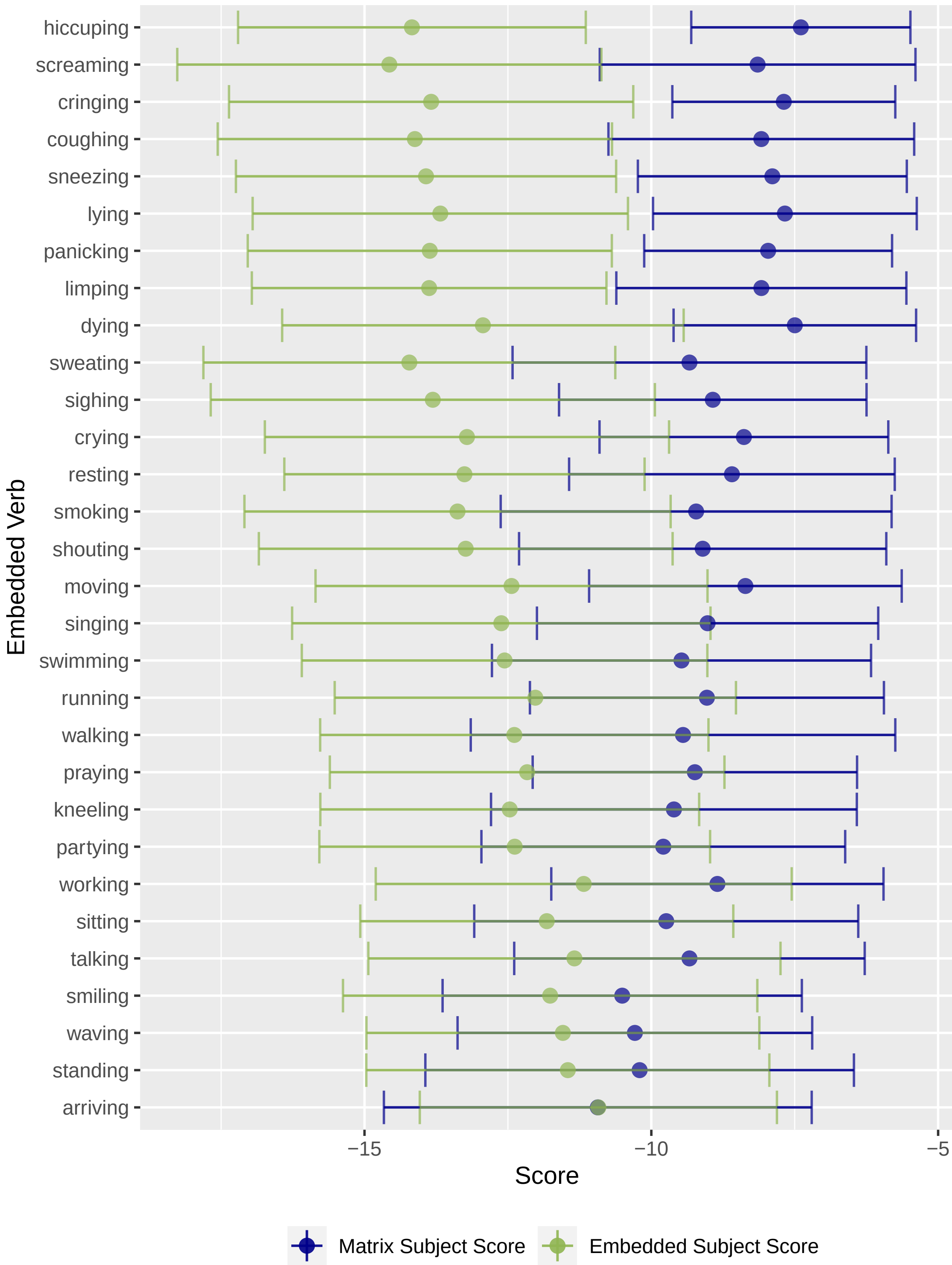


Figure 9: Error bar plot showing mean matrix subject score and embedded subject score for stimuli with each embedded verb. Rows are ordered by matrix subject bias. Error bars show standard deviation.

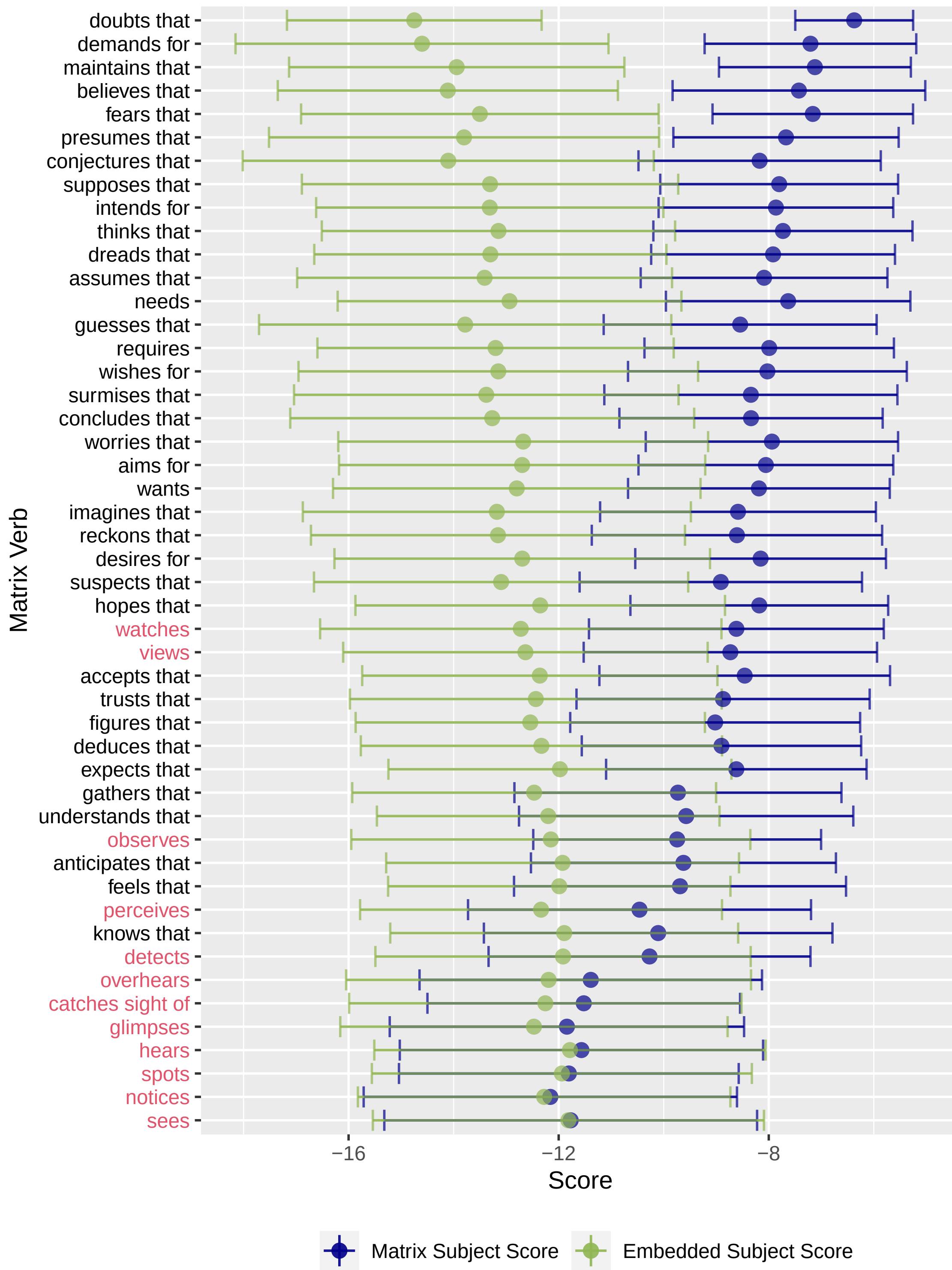


Figure 10: Error bar plot showing mean matrix subject score and embedded subject score for stimuli with each matrix verb. Rows are ordered by matrix subject bias. Error bars show standard deviation. Perceptual matrix verbs are highlighted in red.