Uncovering Constraint-Based Behavior in Neural Models via Targeted Fine-Tuning

Forrest Davis and **Marten van Schijndel**
Department of Linguistics
Cornell University
{fd252|mv443}@cornell.edu

Abstract

A growing body of literature has focused on detailing the linguistic knowledge embedded in large, pretrained language models. Existing work has shown that non-linguistic biases in models can drive model behavior away from linguistic generalizations. We hypothesized that competing linguistic processes within a language, rather than just non-linguistic model biases, could obscure underlying linguistic knowledge. We tested this claim by exploring a single phenomenon in four languages: English, Chinese, Spanish, and Italian. While human behavior has been found to be similar across languages, we find cross-linguistic variation in model behavior. We show that competing processes in a language act as constraints on model behavior and demonstrate that targeted fine-tuning can re-weight the learned constraints, uncovering otherwise dormant linguistic knowledge in models. Our results suggest that models need to learn both the linguistic constraints in a language and their relative ranking, with mismatches in either producing non-human-like behavior.

1 Introduction

Ever larger pretrained language models continue to demonstrate success on a variety of NLP benchmarks (e.g., Devlin et al., 2019; Brown et al., 2020). One common approach for understanding why these models are successful is centered on inferring what linguistic knowledge such models acquire (e.g., Linzen et al., 2016; Hewitt and Manning, 2019; Hu et al., 2020; Warstadt et al., 2020a). Linguistic knowledge alone, of course, does not fully account for model behavior; non-linguistic heuristics have also been shown to drive model behavior (e.g., sentence length; see McCoy et al., 2019; Warstadt et al., 2020b). Nevertheless, when looking across a variety of experimental methods, models appear to acquire some grammatical knowledge (see Warstadt et al., 2019).

However, investigations of linguistic knowledge in language models are limited by the overwhelming prominence of work solely on English (though see Gulordava et al., 2018; Ravfogel et al., 2018; Mueller et al., 2020). Prior work has shown non-linguistic biases of neural language models mimic English-like linguistic structure, limiting the generalizability of claims founded on English data (e.g., Dyer et al., 2019; Davis and van Schijndel, 2020b). In the present study, we show via cross-linguistic comparison, that knowledge of competing linguistic constraints can obscure underlying linguistic knowledge.

Our investigation is centered on a single discourse phenomena, implicit causality (IC) verbs, in four languages: English, Chinese, Spanish, and Italian. When an IC verb occurs in a sentence, interpretations of pronouns are affected:

(1) a. Lavender frightened Kate because she was so terrifying.
b. Lavender admired Kate because she was so amazing.

In (1), both *Lavender* and *Kate* agree in gender with *she*, so both are possible antecedents. However, English speakers overwhelmingly interpret *she* as referring to *Lavender* in (1-a) and *Kate* in (1-b). Verbs that have a subject preference (e.g., *frightened*) are called subject-biased IC verbs, and verbs with an object preference (e.g., *admired*) are called object-biased IC verbs.

IC has been a rich source of psycholinguistic investigation (e.g., Garvey and Caramazza, 1974; Hartshorne, 2014; Williams, 2020). Current accounts of IC ground the phenomenon within the linguistic signal without the need for additional pragmatic inferences by comprehenders (e.g., Ro-

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 1159–1171
August 1–6, 2021. ©2021 Association for Computational Linguistics

hde et al., 2011; Hartshorne et al., 2013). Recent investigations of IC in neural language models confirms that the IC bias of English is learnable, at least to some degree, from text data alone (Davis and van Schijndel, 2020a; Upadhye et al., 2020). The ability of models trained on other languages to acquire an IC bias, however, has not been explored. Within the psycholinguistic literature, IC has been shown to be remarkably consistent cross-linguistically (see Hartshorne et al., 2013; Ngo and Kaiser, 2020). That is, IC verbs have been attested in a variety of languages. Given the cross-linguistic consistency of IC, then, models trained on other languages should also demonstrate an IC bias. However, using two popular model types, BERT based (Devlin et al., 2019) and RoBERTa based (Liu et al., 2019),[1] we find that models only acquired a human-like IC bias in English and Chinese but not in Spanish and Italian.

We relate this to a crucial difference in the presence of a competing linguistic constraint affecting pronouns in the target languages. Namely, Spanish and Italian have a well studied process called *pro drop*, which allows for subjects to be ‘empty’ (Rizzi, 1986). An English equivalent would be “(she) likes BERT” where *she* can be elided. While IC verbs increase the probability of a pronoun that refers to a particular antecedent, pro drop disprefers any overt pronoun in subject position (i.e. the target location in our study). That is, both processes are in direct competition in our experiments. As a result, Spanish and Italian models are susceptible to overgeneralizing any learned pro-drop knowledge, favoring no pronouns rather than IC-conditioned pronoun generation.

To exhibit an IC bias, models of Spanish and Italian have two tasks: learn the relevant constraints (i.e. IC and pro drop) and the relative ranking of these constraints. We find that the models learn both constraints, but, critically, instantiate the wrong ranking, favoring pro drop to an IC bias. Using fine-tuning to demote pro drop, we are able to uncover otherwise dormant IC knowledge in Spanish and Italian. Thus, the apparent failure of the Spanish and Italian models to pattern like English and Chinese is not evidence on its own of a model’s inability to acquire the requisite linguistic knowledge, but is in fact evidence that models are unable to adjudicate between competing linguistic constraints in a human-like way. In English and Chinese, the promotion of a pro-drop process via fine-tuning has the opposing effect, diminishing an IC bias in model behavior. As such, our results indicate that non-human like behavior can be driven by failure either to learn the underlying linguistic constraints or to learn the relevant constraint ranking.

2 Related Work

This work is intimately related to the growing body of literature investigating linguistic knowledge in large, pretrained models. Largely, this literature articulates model knowledge via isolated linguistic phenomena, such as subject-verb agreement (e.g., Linzen et al., 2016; Mueller et al., 2020), negative polarity items (e.g., Marvin and Linzen, 2018; Warstadt et al., 2019), and discourse and pragmatic structure (including implicit causality; e.g., Ettinger, 2020; Schuster et al., 2020; Jeretic et al., 2020; Upadhye et al., 2020). Our study differs, largely, in framing model linguistic knowledge as sets of competing constraints, which privileges the interaction between linguistic phenomena.

Prior work has noted competing generalizations influencing model behavior via the distinction of non-linguistic vs. linguistic biases (e.g., McCoy et al., 2019; Davis and van Schijndel, 2020a; Warstadt et al., 2020b). The findings in Warstadt et al. (2020b), that linguistic knowledge is represented within a model much earlier than attestation in model behavior, bears resemblance to our claims. We find that linguistic knowledge can, in fact, lie dormant due to other linguistic processes in a language, not just due to non-linguistic preferences. Our findings suggest that some linguistic knowledge may never surface in model behavior, though further work is needed on this point.

In the construction of our experiments, we were inspired by synthetic language studies which probe the underlying linguistic capabilities of language models (e.g., McCoy et al., 2018; Ravfogel et al., 2019). We made use of synthetically modified language data that accentuated, or weakened, evidence for certain linguistic processes. The goal of such modification in our work is quite similar both to work which attempts to remove targeted linguistic knowledge in model representations (e.g., Ravfogel et al., 2020; Elazar et al., 2021) and to work which

[1] These model types were chosen for ease of access to existing models. Pretrained, large auto-regressive models are largely restricted to English, and prior work suggests that LSTMs are limited in their ability to acquire an IC bias in English (Davis and van Schijndel, 2020a).

Model	Lang	Tokens
BERT	EN	3.3B
RoBERTa	EN	30B
Chinese BERT	ZH	5.4B
Chinese RoBERTa	ZH	5.4B
BETO	ES	3B
RuPERTa	ES	3B
Italian BERT	IT	2B
UmBERTo	IT	0.6B
GilBERTo	IT	11B

Table 1: Summary of models investigated with language and approximate number of tokens in training. For RoBERTa we use the approximation given in Warstadt et al. (2020b).

investigates the representational space of models via priming (Prasad et al., 2019; Misra et al., 2020). In the present study, rather than identifying isolated linguistic knowledge or using priming to study relations between underlying linguistic representations, we ask **how linguistic representations interact to drive model behavior**.

3 Models

Prior work on IC in neural language models has been restricted to autoregressive models for ease of comparison to human results (e.g., Upadhye et al., 2020). In the present study, we focused on two popular non-autoregressive language model variants, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). We used existing models available via HuggingFace (Wolf et al., 2020).

Multilingual models have been claimed to perform worse on targeted linguistics tasks than monolingual models (e.g., Mueller et al., 2020). We confirmed this claim by evaluating mBERT which exhibited no IC bias in any language.[2] Thus, we focus in the rest of this paper on monolingual models (summarized in Table 1). For English, we used the BERT base uncased model and the RoBERTa base model. For Chinese, we evaluated BERT and RoBERTa models from Cui et al. (2020). For Spanish, we used BETO (Cañete et al., 2020) and RuPERTa (Romero, 2020). For Italian, we evaluated an uncased Italian BERT [3] as well as two RoBERTa based models, UmBERTo (Parisi et al., 2020) and GilBERTo (Ravasio and Di Perna, 2020).

4 Experimental Stimuli and Measures

Our list of target verbs was derived from existing psycholinguistic studies of IC verbs.[4] For English, we used the IC verbs from Ferstl et al. (2011).

Each verb in the human experiment was coded for IC bias based on continuations of sentence fragments (e.g., *Kate accused Bill because* ...). For Spanish, we used the IC verbs from Goikoetxea et al. (2008), which followed a similar paradigm as Ferstl et al. (2011) for English. Participants were given sentence fragments and asked to complete the sentence and circle their intended referent. The study reported the percent of subject continuations for 100 verbs, from which we used the 61 verbs which had a significant IC bias (i.e. excluding verbs with no significant subject or object bias).

For Italian, we used the 40 IC verbs reported in Mannetti and De Grada (1991). Human participants were given ambiguous completed sentences with no overt pronoun like "John feared Michael because of the kind of person (he) is" and were asked to judge who the null pronoun referred to, with the average number of responses that gave the subject as the antecedent reported.[5] For Chinese, we used 59 IC verbs reported in Hartshorne et al. (2013), which determined average subject bias per verb in a similar way as Mannetti and De Grada (1991) (i.e. judgments of antecedent preferences given ambiguous sentences, this time with overt pronouns).[6]

We generated stimuli using 14 pairs of stereotypical male and female nouns (e.g., *man* vs. *woman*, *husband* vs. *wife*) in each language, rather than rely on proper names as was done in the human experiments. The models we investigated are bidirectional, so we used a neutral right context, *was there*, for English and Spanish, where human ex-

[2] Results are provided in Appendix B

[3] https://huggingface.co/dbmdz/bert-base-italian-uncased

[4] All stimuli, as well as code for reproducing the results of the paper are available at `https://github.com/forrestdavis/ImplicitCausality`. For each language investigated, the stimuli were evaluated for grammaticality by native speakers with academic training in linguistics.

[5] Specifically, Mannetti and De Grada (1991) grouped the verbs into four categories and reported the average per category as well as individual verb results for the most biased verbs and the negative/positive valency verbs. Additionally, figures showing average responses across various conditions was reported for one of the categories. From the combination of this information, the average scores for all but two verbs were able to be determined. The remaining two verbs were assigned the reported average score of their stimuli group.

[6] In Hartshorne et al. (2013), 60 verbs were reported, but after consultation with a native speaker with academic training in linguistics, one verb was excluded due to perceived ungrammaticality of the construction.

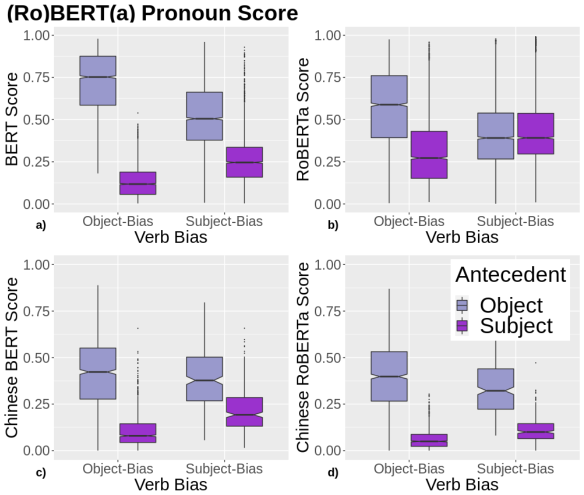

Figure 1: Model scores for **a)** BERT, **b)** RoBERTa, **c)** Chinese BERT, and **d)** Chinese RoBERTa at the pronoun grouped by antecedent; stimuli derived from Ferstl et al. (2011) and Hartshorne et al. (2013)

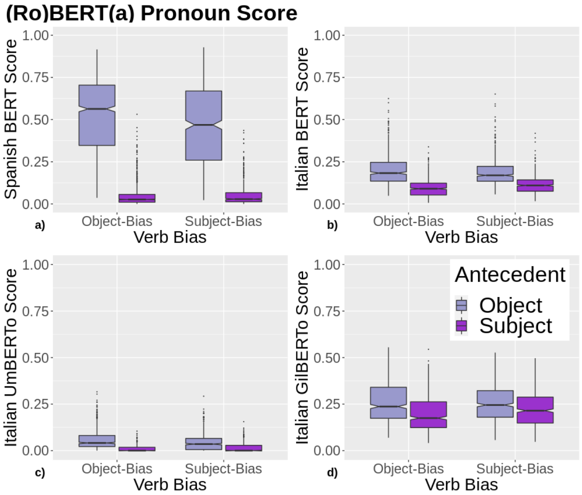

Figure 2: Model scores for **a)** Spanish BERT (BETO), **b)** Italian BERT, **c)** UmBERTo, and **d)** GilBERTo at the pronoun grouped by antecedent; stimuli derived from Goikoetxea et al. (2008) and Mannetti and De Grada (1991)

periments provided no right context.[7] For Italian we utilized the full sentences investigated in the human experiments. The Chinese human experiment also used full sentences, but relied on nonce words (i.e. novel, constructed words like sliktopoz), so we chose instead to generate sentences like the English and Spanish ones. All stimuli had subjects and objects that differed in gender, such that all nouns occurred in subject or object position (i.e. the stimuli were fully balanced for gender):

(2) the man admired the woman because [MASK] was there.[8]

The mismatch in gender forced the choice of pronoun to be unambiguous. For each stimulus, we gathered the scores assigned to the third person singular male and female pronouns (e.g., *he* and *she*).[9] Our measures were grouped by antecedent type (i.e. the pronoun refers to the subject or the object) and whether the verb was object-biased or subject-biased. For example, BERT assigns to (2) a score of 0.01 for the subject antecedent (i.e. *he*) and 0.97 for the object (i.e. *she*), in line with the object-bias of *admire*.

5 Models Inconsistently Capture Implicit Causality

As exemplified in (1), repeated below, IC verb bias modulates the preference for pronouns.

(3) a. Lavender frightened Kate because she was so terrifying.
b. Lavender admired Kate because she was so amazing.

An object-biased IC verb (e.g., *admired*) should increase the likelihood of pronouns that refer to the object, and a subject-biased IC verb (e.g., *frightened*) should increase the likelihood of reference to the subject. Given that all the investigated stimuli were disambiguated by gender, we categorized our results by the antecedent of the pronoun and the IC verb bias. We first turn to English and Chinese, which showed an IC bias in line with existing work on IC bias in autoregressive English models (e.g., Upadhye et al., 2020; Davis and van Schijndel, 2020a). We then detail the results for Spanish and Italian, where only very limited, if any, IC bias was observed.

5.1 English and Chinese

The results for English and Chinese are given in Figure 1 and detailed in Appendix B. All models demonstrated a greater preference for pronouns referring to the object after an object-biased IC verb

[7] Using *here, outside,* or *inside* as the right context produces qualitatively the same patterns.

[8] The model-specific mask token was used. Additionally, all models were uncased, with the exception of RoBERTa, so lower cased stimuli were used.

[9] In spoken Chinese, the male and female pronouns are homophonous. They are, however, distinguished in writing.

than after a subject-biased IC verb.[10] Additionally, they had greater preferences for pronouns referring to the subject after a subject-biased IC verb than after a object-biased IC verb. That is, all models showed the expected IC-bias effect. Generally, there was an overall greater preference for referring to the object, in line with a recency bias, with the exception of RoBERTa, where subject-biased IC verbs neutralized the recency effect.

5.2 Spanish and Italian

The results for Spanish and Italian are given in Figure 2 and detailed in Appendix B. In stark contrast to the models of English and Chinese, an IC bias was either not demonstrated or was only weakly attested. For Spanish, BETO showed a greater preference for pronouns referencing the object after an object-biased IC verb than after a subject-biased IC verb. There was no corresponding IC effect for pronouns referring to the subject, and RuPERTa (a RoBERTa based model) had no IC effect at all.

Italian BERT and GilBERTo (a RoBERTa based model) had no significant effect of IC-verb on pronouns referring to the object. There was a significant, albeit very small, increased score for pronouns referring to the subject after a subject-biased IC verb in line with a weak subject-IC bias. Similarly, UmBERTo (a RoBERTa based model) had significant, yet tiny IC effects, where object-biased IC verbs increased the score of pronouns referring to objects compared to subject-biased IC verbs (conversely with pronouns referring to the subject).

Any significant effects in Spanish and Italian were much smaller than their counterparts in English (as is visually apparent between Figure 1 and Figure 2), and each of the Spanish and Italian models failed to demonstrate at least one of the IC effects.

6 Pro Drop and Implicit Causality: Competing Constraints

We were left with an apparent mismatch between models of English and Chinese and models of Spanish and Italian. In the former, an IC verb bias modulated pronoun preferences. In the latter, the same IC verb bias was comparably absent. Recall that, for humans, the psycholinguistic literature suggests that IC bias is, in fact, quite consistent across languages (see Hartshorne et al., 2013).

We found a possible reason for why the two sets of models behave so differently by carefully considering the languages under investigation. Languages can be thought of as systems of competing linguistic constraints (e.g., Optimality Theory; Prince and Smolensky, 2004). Spanish and Italian exhibit pro drop and typical grammatical sentences often lack overt pronouns in subject position, opting instead to rely on rich agreement systems to disambiguate the intended subject at the verb (Rizzi, 1986). This constraint competes with IC, which favors pronouns that refer to either the subject or the object. Chinese also allows for empty arguments (both subjects and objects), typically called *discourse pro-drop* (Huang, 1984).[11] As the name suggests, however, this process is more discourse constrained than the process in Spanish and Italian. For example, in Chinese, the empty subject can only refer to the subject of the preceding sentence (see Liu, 2014). As a means of comparison, in surveying three Universal Dependencies datasets,[12] 8% of nsubj (or nsubj:pass) relations were pronouns for Chinese, while only 2% and 3% were pronouns in Spanish and Italian respectively. English lies on the opposite end of the continuum, requiring overt pronouns in the absence of other nominals (cf. *He likes NLP* and **Likes NLP*).

Therefore, it's possible that the presence of competing constraints in Spanish and Italian obscured the underlying IC knowledge: one constraint preferring pronouns which referred to the subject or object and the other constraint penalizing overt pronouns in subject positions (i.e. the target position masked in our experiments). In the following sections, we removed or otherwise demoted the dominance of each model's pro-drop constraint for Spanish and Italian, and introduced or promoted a pro-drop like constraint in English and Chinese. We found that the degree of IC bias in model behavior could be controlled by the presence, or absence, of a competing pro-drop constraint.

6.1 Methodology

We constructed two classes of dataset to fine-tune the models on. The first aimed to demote the pro-

[10] Throughout the paper, statistical significance was determined by two-way *t*-tests evaluating the difference between pronouns referring to objects after subject-biased and object-biased IC verbs, and similarly for pronouns referring to the subject. The threshold for statistical significance was p = 0.0009, after adjusting for the 54 statistical tests conducted in the paper.

[11] Other names common to the literature include *topic drop*, *radical pro drop*, and *rampant pro drop*.

[12] Chinese GSD, Italian ISDT, and Spanish AnCora.

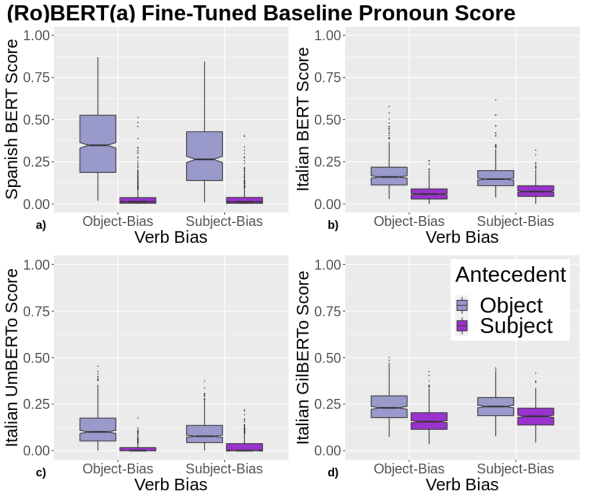

Figure 3: After fine-tuning on baseline data (i.e. pro-drop sentences), model scores for **a)** Spanish BERT (BETO), **b)** Italian BERT, **c)** UmBERTo, and **d)** GilBERTo at the pronoun grouped by antecedent; stimuli derived from Goikoetxea et al. (2008) and Mannetti and De Grada (1991)

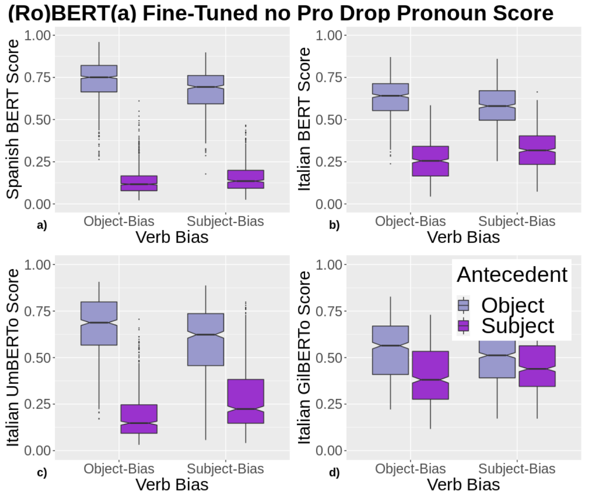

Figure 4: After fine-tuning on sentences removing pro drop (i.e. adding a subject pronoun), model scores for **a)** Spanish BERT (BETO), **b)** Italian BERT, **c)** UmBERTo, and **d)** GilBERTo at the pronoun grouped by antecedent; stimuli derived from Goikoetxea et al. (2008) and Mannetti and De Grada (1991)

drop constraint in Spanish and Italian. The second aimed to inject a pro-drop constraint into English and Chinese. For both we relied on Universal Dependencies datasets. For Spanish, we used the AnCora Spanish newswire corpus (Taulé et al., 2008), for Italian we used ISDT (Bosco et al., 2013) and VIT (Delmonte et al., 2007), for English we used the English Web Treebank (Silveira et al., 2014), and for Chinese, we used the Traditional Chinese Universal Dependencies Treebank annotated by Google (GSD) and the Chinese Parallel Universal Dependencies (PUD) corpus from the 2017 CoNLL shared task (Zeman et al., 2017).

For demoting pro drop, we found finite (i.e. inflected) verbs that did not have a subject relation in the corpora.[13] We then added a pronoun, matching the person and number information given on the verb, alternating the gender. For Italian, this amounted to a dataset of 3798 sentences with a total of 4608 pronouns (2,284 he or she) added. For parity with Italian, we restricted Spanish to a dataset of the first 4000 sentences, which had 5,559 pronouns (3,573 he or she) added. For the addition of a pro-drop constraint in English and Chinese, we found and removed pronouns that bore a subject relation to a verb. This amounted to 935 modified sentences and 1083 removed pronouns (774 he or she) in Chinese and 4000 modified sentences and 5984 removed pronouns (2188 he or she) in English.[14]

For each language, 500 unmodified sentences were used for validation, and unchanged versions of all the sentences were kept and used to fine-tune the models as a baseline to ensure that there was nothing about the data themselves that changed the IC-bias of the models. Moreover, the fine-tuning data was filtered to ensure that no verbs evaluated in our test data were included. Fine-tuning proceeded using HuggingFace's API. Each model was fine-tuned with a masked language modeling objective for 3 epochs with a learning rate of 5e-5, following the fine-tuning details in (Devlin et al., 2019).[15]

6.2 Demoting Pro Drop: Spanish and Italian

As a baseline, we fine-tuned the Spanish and Italian models on unmodified versions of all the data we used for demoting pro drop. The baseline results are given in Figure 3. We found the same qualitative effects detailed in Section 5.2, confirming that the data used for fine-tuning on their own did not produce model behavior in line with an IC bias.

We turn now to our main experimental manipu-

[13] In particular, verbs that lacked any nsubj, nsubj:pass, expl, expl:impers, or expl:pass dependents

[14] A fuller breakdown of the fine-tuning data is given in Appendix A with the full training and evaluation data given on our Github. We restricted English to the first 4000 sentences for parity with Italian/Spanish. Using the full set of sentences resulted in qualitatively the same pattern. We used the maximum number of sentences we could take from Chinese UD.

[15] We provide a Colab script for reproducing all fine-tuned models on our Github.

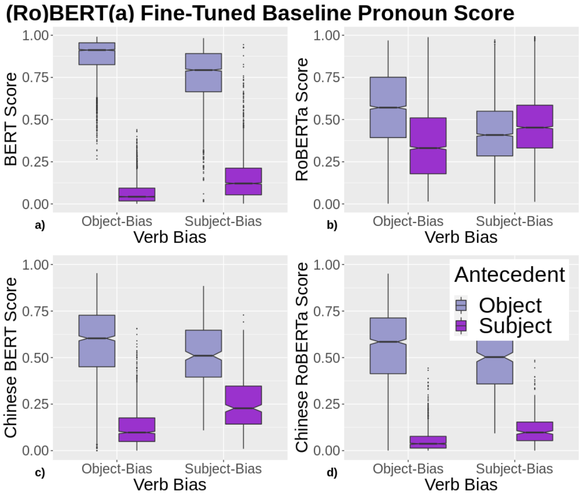

Figure 5: After fine-tuning on baseline data (i.e. without removing subject pronouns), model scores for **a)** BERT, **b)** RoBERTa, **c)** Chinese BERT, and **d)** Chinese RoBERTa at the pronoun grouped by antecedent; stimuli derived from Ferstl et al. (2011) and Hartshorne et al. (2013)

lation: fine-tuning the Spanish and Italian models on sentences that exhibit the opposite of a pro-drop effect. It is worth repeating that the fine-tuning data shared no verbs or sentence frames with our test data. The results are given in Figure 4. Strikingly, an object-biased IC effect (pronouns referring to the object were more likely after object-biased IC verbs than subject-biased IC verbs) was observed for Italian BERT and GilBERTo despite no such effect being observed in the base models. Moreover, both models showed a more than doubled subject-biased IC verb effect. UmBERTo also showed increased IC effects, as compared to the base models. Similarly for Spanish, a subject-biased IC verb effect materialized for BETO when no corresponding effect was observed with the base model. The object-biased IC verb effect remained similar to what was reported in Section 5.2. For RuPERTa, which showed no IC knowledge in the initial investigation, no IC knowledge surfaced after fine-tuning. We conclude that RuPERTa has no underlying knowledge of IC, though further work should investigate this claim.

Taken together these results indicate that simply fine-tuning on a small number of sentences can re-rank the linguistic constraints influencing model behavior and uncover other linguistic knowledge (in our case an underlying IC-bias). That is, model behavior can hide linguistic knowledge not just because of non-linguistic heuristics, but also due to over-zealously learning one isolated aspect of linguistic structure at the expense of another.

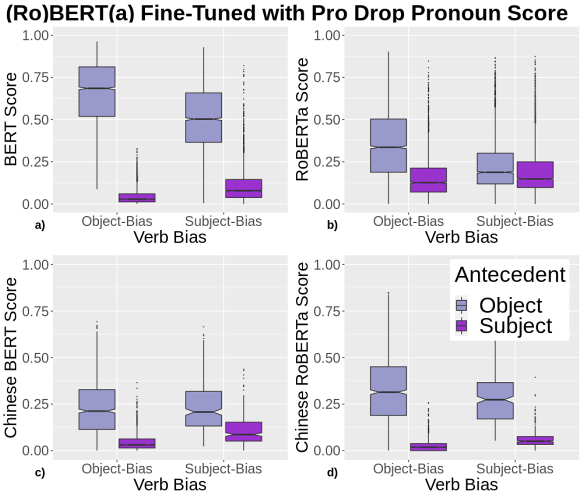

Figure 6: After fine-tuning on sentences with pro drop (i.e. no subject pronouns), model scores for **a)** BERT, **b)** RoBERTa, **c)** Chinese BERT, and **d)** Chinese RoBERTa at the pronoun grouped by antecedent; stimuli derived from Ferstl et al. (2011) and Hartshorne et al. (2013)

6.3 Promoting Pro Drop: English and Chinese

Next, we fine-tune a pro-drop constraint into models of English and Chinese. Recall that both models showed an IC effect, for both object-biased and subject-biased IC verbs. Moreover, both languages lack the pro-drop process found in Spanish and Italian (though Chinese allows null arguments).

As with Spanish and Italian, we fine-tuned the English and Chinese models on unmodified versions of the training sentences as a baseline (i.e. the sentences kept their pronouns) with the results given in Figure 5. There was no qualitative difference from the IC effects noted in Section 5.1. That is, for both English and Chinese, pronouns referring to the object were more likely after object-biased IC verbs than after subject-biased IC verbs, and conversely pronouns referring to the subject were more likely after subject-biased than object-biased IC verbs.

The results after fine-tuning the models on data mimicking a Spanish and Italian like pro-drop process (i.e. no pronouns in subject position) are given in Figure 6 and detailed in Appendix B. Despite fine-tuning on only 0.0004% and 0.003% of the data RoBERTa and BERT were trained on, respectively, the IC effects observed in Section 5.1 were severely diminished in English. However, the subject-biased IC verb effect remained robust in both models. For Chinese BERT, the subject-biased IC verb effect in the base model was lost and the object-biased IC verb effect was reduced. The subject-biased IC verb effect was similarly attenuated in Chinese RoBERTa. However, the object-biased IC verb effect remained.

For both languages, exposure to relatively little pro-drop data weakened the IC effect in behavior and even removed it in the case of subject-biased IC verbs in Chinese BERT. This result strengthens our claim that competition between learned linguistic constraints can obscure underlying linguistic knowledge in model behavior.

7 Discussion

The present study investigated the ability of RoBERTa and BERT models to demonstrate knowledge of implicit causality across four languages (recall the contrast between *Lavender frightened Kate* and *Lavender admired Kate* in (1)). Contrary to humans, who show consistent subject and object-biased IC verb preferences across languages (see Hartshorne et al., 2013), BERT and RoBERTa models of Spanish and Italian failed to demonstrate the full IC bias found in English and Chinese BERT and RoBERTa models (with our English results supporting prior work on IC bias in neural models and extending it to non-autoregressive models; Upadhye et al., 2020; Davis and van Schijndel, 2020a). Following standard behavioral probing (e.g., Linzen et al., 2016), this mismatch could be interpreted as evidence of differences in linguistic knowledge across languages. That is, model behavior in Spanish and Italian was inconsistent with predictions from the psycholinguistic IC literature, suggesting that these models lack knowledge of implicit causality. However, we found that to be an incorrect inference; the models *did* have underlying knowledge of IC.

Other linguistic processes influence pronouns in Spanish and Italian, and we showed that competition between multiple distinct constraints affects model behavior. One constraint (pro drop) decreases the probability of overt pronouns in subject position, while the other (IC) increases the probability of pronouns that refer to particular antecedents (subject-biased verbs like *frightened* favoring subjects and object-biased verbs like *admired* favoring objects). Models of Spanish and Italian, then, must learn not only these two constraints, but also their ranking (i.e. should the model generate a pronoun as IC dictates, or generate no pronoun in line with pro drop). By fine-tuning the models on data contrary to pro drop (i.e. with overt pronouns in subject position), we uncovered otherwise hidden IC knowledge. Moreover, we found that fine-tuning a pro-drop constraint into English and Chinese greatly diminished IC's influence on model behavior (with as little as 0.0004% of a models original training data).

Taken together, we conclude that there are two ways of understanding mismatches between model linguistic behavior and human linguistic behavior. Either a model fails to learn the necessary linguistic constraint, or it succeeds in learning the constraint but fails to learn the correct interaction with other constraints. Existing literature points to a number of reasons a model may be unable to learn a linguistic representation, including the inability to learn mappings between form and meaning and the lack of embodiment (e.g., Bender and Koller, 2020; Bisk et al., 2020). We suggest that researchers should re-conceptualize linguistic inference on the part of neural models as inference of constraints and constraint ranking in order to better understand model behavior. We believe such framing will open additional connections with linguistic theory and psycholinguistics. Minimally, we believe targeted fine-tuning for constraint re-ranking may provide a general method both to understand what linguistic knowledge these models possess and to aid in making their linguistic behavior more human-like.

8 Conclusion and Future Work

The present study provided evidence that model behavior can be meaningfully described, and understood, with reference to competing constraints. We believe that this is a potentially fruitful way of reasoning about model linguistic knowledge. Possible future directions include pairing our behavioral analyses with representational probing in order to more explicitly link model representations and model behavior (e.g., Ettinger et al., 2016; Hewitt and Liang, 2019) or exploring constraint competition in different models, like GPT-2 which has received considerable attention for its apparent linguistic behavior (e.g., Hu et al., 2020) and its ability to predict neural responses (e.g., Schrimpf et al., 2020).

Acknowledgments

We would like to thank members of the C.Psyd Lab, the Cornell NLP group, and the Stanford NLP Group, who gave valuable feedback on earlier forms of this work. Thanks also to the anonymous reviewers whose comments improved the paper.

References

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.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8718–8735, Online. Association for Computational Linguistics.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian treebanks: Towards an Italian Stanford dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69, Sofia, Bulgaria. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Forrest Davis and Marten van Schijndel. 2020a. 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.

Forrest Davis and Marten van Schijndel. 2020b. Recurrent Neural Network Language Models Always Learn English-Like Relative Clause Attachment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1979–1990, Online. Association for Computational Linguistics.

Rodolfo Delmonte, Antonella Bristot, and Sara Tonelli. 2007. VIT – Venice Italian Treebank: Syntactic and Quantitative Features. In *Sixth International Workshop on Treebanks and Linguistic Theories*, volume 1, pages 43–54.

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.

Chris Dyer, Gábor Melis, and Phil Blunsom. 2019. A Critical Analysis of Biased Parsers in Unsupervised Parsing. *arXiv:1909.09428 [cs]*.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

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, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139. Association for Computational Linguistics.

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.

Catherine Garvey and Alfonso Caramazza. 1974. Implicit Causality in Verbs. *Linguistic Inquiry*, 5(3):459–464.

Edurne Goikoetxea, Gema Pascual, and Joana Acha. 2008. Normative study of the implicit causality of 100 interpersonal verbs in Spanish. *Behavior Research Methods*, 40(3):760–772.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless Green Recurrent Networks Dream Hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205. Association for Computational Linguistics.

Joshua K. Hartshorne. 2014. What is implicit causality? *Language, Cognition and Neuroscience*, 29(7):804–824.

Joshua K. Hartshorne, Yasutada Sudo, and Miki Uruwashi. 2013. Are implicit causality pronoun resolution biases consistent across languages and cultures? *Experimental Psychology*, 60(3):179–196.

John Hewitt and Percy Liang. 2019. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2733–2743. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4129–4138. Association for Computational Linguistics.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A Systematic Assessment of Syntactic Generalization in Neural Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

C.-T. James Huang. 1984. On the Distribution and Reference of Empty Pronouns. *Linguistic Inquiry*, 15(4):531–574.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4(0):521–535.

Chi-Ming Louis Liu. 2014. *A Modular Theory of Radical Pro Drop*. Ph.D., Harvard University.

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:1907.11692 [cs]*.

L. Mannetti and E. De Grada. 1991. Interpersonal verbs: Implicit causality of action verbs and contextual factors: Implicit causality of action verbs. *European Journal of Social Psychology*, 21(5):429–443.

Rebecca Marvin and Tal Linzen. 2018. Targeted Syntactic Evaluation of Language Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202. Association for Computational Linguistics.

R Thomas McCoy, Robert Frank, and Tal Linzen. 2018. Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks. *Proceedings of the 40th Annual Virtual Meeting of the Cognitive Science Society*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. Exploring BERT's Sensitivity to Lexical Cues using Tests from Semantic Priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635, Online. Association for Computational Linguistics.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-Linguistic Syntactic Evaluation of Word Prediction Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539. Association for Computational Linguistics.

Binh Ngo and Elsi Kaiser. 2020. Implicit Causality: A Comparison of English and Vietnamese Verbs. *University of Pennsylvania Working Papers in Linguistics*, 26.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. UmBERTo: An Italian Language Model trained with Whole Word Masking.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. 2019. Using Priming to Uncover the Organization of Syntactic Representations in Neural Language Models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 66–76. Association for Computational Linguistics.

Alan Prince and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell Pub., Malden, MA.

Giulio Ravasio and Leonardo Di Perna. 2020. GilBERTo: An Italian pretrained language model based on RoBERTa.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. 2019. Studying the Inductive Biases of RNNs with Synthetic Variations of Natural Languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3532–3542. Association for Computational Linguistics.

Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. 2018. Can LSTM Learn to Capture Agreement? The Case of Basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107. Association for Computational Linguistics.

Luigi Rizzi. 1986. Null Objects in Italian and the Theory of pro. *Linguistic Inquiry*, 17(3):501–557.

H. Rohde, R. Levy, and A. Kehler. 2011. Anticipating explanations in relative clause processing. *Cognition*, 118(3):339–358.

Manuel Romero. 2020. RuPERTa: The Spanish RoBERTa.

Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. 2020. Artificial Neural Networks Accurately Predict Language Processing in the Brain. *bioRxiv*, page 2020.06.26.174482.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5387–5403, Online. Association for Computational Linguistics.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for catalan and spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.

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.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020a. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020b. Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 217–235, Online. Association for Computational Linguistics.

Elyce Dominique Williams. 2020. *Language Experience Predicts Pronoun Comprehension in Implicit Causality Sentences*. Master's, University of North Carolina at Chapel Hill.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19. Association for Computational Linguistics.

A Additional Fine-tuning Training Information

The full breakdown of pronouns added or removed in the fine-tuning training data are detailed below. English can be found in Table 2, Chinese can be found in Table 3, Spanish can be found in Table 4, and Italian can be found in Table 5.

	SG	PL	NA
1	1927	617	-
2	-	-	1252
3	1548	640	-

Table 2: Breakdown of pronouns removed for English fine-tuning data. Pronoun person and number were determined by annotations in UD data, with NA being pronouns unmarked for number. There were a total of 4000 sentences comprised of 66929 tokens in the training set.

	SG	PL	NA
1	-	56	66
2	-	2	21
3	-	164	774

Table 3: Breakdown of pronouns removed for Chinese fine-tuning data. Pronoun person and number were determined by annotations in UD data, with NA being pronouns unmarked for number. There were a total of 935 sentences comprised of 108949 characters in the training set.

	SG	PL	NA
1	519	417	-
2	99	7	-
3	3574	944	-

Table 4: Breakdown of pronouns added for Spanish fine-tuning data. Pronoun person and number were determined by annotations in UD data, with NA being pronouns unmarked for number. There were a total of 4000 sentences comprised of 5559 tokens in the training set.

	SG	PL	NA
1	654	417	-
2	399	94	-
3	2284	679	-

Table 5: Breakdown of pronouns added for Italian fine-tuning data. Pronoun person and number were determined by annotations in UD data, with NA being pronouns unmarked for number. There were a total of 3798 sentences comprised of 4608 tokens in the training set.

B Expanded Results (including mBERT)

The full details of the pairwise *t*-tests conducted for the present study are given below (including the results for mBERT). The results for English models are in Table 6, for Chinese models Table 7, for Spanish models Table 8, and Italian models Table 9.

model	O-O μ	O-S μ	CI	p	S-O μ	S-S μ	CI	p
BERT	0.72	0.52	[0.19,0.21]	$< 2.2e^{-16}$	0.13	0.26	[0.12,0.13]	$< 2.2e^{-16}$
BERT_BASE	0.75	0.52	[0.11,0.13]	$< 2.2e^{-16}$	0.06	0.15	[0.08,0.09]	$< 2.2e^{-16}$
BERT_PRO	0.51	0.52	[0.14,0.15]	$< 2.2e^{-16}$	0.04	0.11	[0.06,0.07]	$< 2.2e^{-16}$
RoBERTa	0.57	0.41	[0.15,0.17]	$< 2.2e^{-16}$	0.31	0.43	[0.11,0.13]	$< 2.2e^{-16}$
RoBERTa_BASE	0.58	0.45	[0.11,0.13]	$< 2.2e^{-16}$	0.31	0.37	[0.07,0.08]	$< 2.2e^{-16}$
RoBERTa_PRO	0.35	0.23	[0.11,0.13]	$< 2.2e^{-16}$	0.16	0.19	[0.03,0.04]	$< 2.2e^{-16}$
mBERT	0.58	0.59	[-0.003,-0.01]	0.001	0.29	0.28	[-0.002,-0.01]	0.0002

Table 6: Results from pairwise *t*-tests for English across the investigated models. O-O refers to object antecedent after object-biased IC verb and O-S to object antecedent after subject-biased IC verb (similarly for subject antecedents S-O and S-S). CI is 95% confidence intervals (where positive is an IC effect). BERT_BASE and BERT_PRO refer to models fine-tuned on baseline data and data with a pro-drop process respectively.

model	O-O μ	O-S μ	CI	p	S-O μ	S-S μ	CI	p
BERT	0.41	0.39	[0.003,0.05]	0.00003	0.11	0.22	[0.09,0.12]	$< 2.2e^{-16}$
BERT_BASE	0.53	0.47	[0.03,0.08]	$2.2e^{-6}$	0.12	0.25	[0.11,0.14]	$< 2.2e^{-16}$
BERT_PRO	0.23	0.23	[-0.02,0.02]	0.94	0.04	0.11	[0.05,0.07]	$< 2.2e^{-16}$
RoBERTa	0.40	0.33	[0.04,0.08]	$1.16e^{-9}$	0.06	0.12	[0.04,0.06]	$< 2.2e^{-16}$
RoBERTa_BASE	0.52	0.46	[0.04,0.08]	$8.4e^{-7}$	0.05	0.11	[0.05,0.07]	$< 2.2e^{-16}$
RoBERTa_PRO	0.32	0.29	[0.002,0.06]	$7e^{-6}$	0.03	0.06	[0.02,0.04]	$< 2.2e^{-16}$
mBERT	0.08	0.07	[0.01,0.03]	$2e^{-6}$	0.08	0.06	[-0.009,-0.002]	$1.3e^{-5}$

Table 7: Results from pairwise *t*-tests for Chinese across the investigated models. O-O refers to object antecedent after object-biased IC verb and O-S to object antecedent after subject-biased IC verb (similarly for subject antecedents S-O and S-S). CI is 95% confidence intervals (where positive is an IC effect). BERT_BASE and BERT_PRO refer to models fine-tuned on baseline data and data with a pro-drop process respectively.

model	O-O μ	O-S μ	CI	p	S-O μ	S-S μ	CI	p
BERT	0.53	0.46	[0.04,0.09]	$1.4e^{-8}$	0.05	0.05	[0.0007,0.01]	0.03
BERT_BASE	0.37	0.30	[0.05,0.08]	$8e^{-12}$	0.03	0.03	[-0.004,0.007]	0.61
BERT_PRO	0.73	0.67	[0.05,0.07]	$< 2.2e^{-16}$	0.16	0.13	[0.01,0.03]	$1.2e^{-7}$
RoBERTa	0.09	0.10	[-0.008,-0.01]	0.03	0.06	0.06	[0.0007,0.007]	0.02
RoBERTa_BASE	0.06	0.06	[-0.005,-0.002]	0.0002	0.04	0.04	[-0.0003,0.004]	0.09
RoBERTa_PRO	0.48	0.48	[-0.03,0.01]	0.42	0.29	0.30	[-0.006,0.02]	0.24
mBERT	0.12	0.11	[0.001,0.01]	0.02	0.02	0.02	[-0.0002,-0.002]	0.03

Table 8: Results from pairwise *t*-tests for Spanish across the investigated models. O-O refers to object antecedent after object-biased IC verb and O-S to object antecedent after subject-biased IC verb (similarly for subject antecedents S-O and S-S). CI is 95% confidence intervals (where positive is an IC effect). BERT_BASE and BERT_PRO refer to models fine-tuned on baseline data and data with a pro-drop process respectively.

model	O-O μ	O-S μ	CI	p	S-O μ	S-S μ	CI	p
BERT	0.21	0.19	[0.005,0.03]	0.004	0.09	0.11	[0.01,0.03]	$1.3e^{-9}$
BERT_BASE	0.17	0.16	[0.006,0.02]	0.002	0.06	0.08	[0.01,0.02]	$4e^{-6}$
BERT_PRO	0.63	0.56	[0.04,0.07]	$1e^{-13}$	0.26	0.32	[0.05,0.07]	$< 2.2e^{-16}$
UmBERTo	0.06	0.05	[0.01,0.02]	$4e^{-6}$	0.009	0.02	[0.004,0.01]	$2e^{-9}$
UmBERTo_BASE	0.12	0.09	[0.02,0.04]	$3e^{-9}$	0.01	0.02	[0.01,0.02]	$9e^{-12}$
UmBERTo_PRO	0.67	0.58	[0.07,0.11]	$5e^{-16}$	0.19	0.28	[0.07,0.11]	$< 2.2e^{-16}$
GilBERTo	0.26	0.25	[-0.006,0.02]	0.30	0.20	0.22	[0.01,0.03]	0.0002
GilBERTo_BASE	0.24	0.24	[-0.006,0.01]	0.44	0.16	0.18	[0.01,0.03]	$3e^{-7}$
GilBERTo_PRO	0.54	0.50	[0.03,0.06]	$3e^{-7}$	0.40	0.45	[0.04,0.07]	$3e^{-10}$
mBERT	0.13	0.14	[-0.004,-0.02]	0.0003	0.12	0.13	[0.003,0.02]	0.003

Table 9: Results from pairwise *t*-tests for Italian across the investigated models. O-O refers to object antecedent after object-biased IC verb and O-S to object antecedent after subject-biased IC verb (similarly for subject antecedents S-O and S-S). CI is 95% confidence intervals (where positive is an IC effect). BERT_BASE and BERT_PRO refer to models fine-tuned on baseline data and data with a pro-drop process respectively.