# A Closer Look at Linguistic Knowledge in Masked Language Models: The Case of Relative Clauses in American English

**Marius Mosbach**\*, **Stefania Degaetano-Ortlieb**\*, **Marie-Pauline Krielke,**
**Badr M. Abdullah**, **Dietrich Klakow**
Department of Language Science and Technology, Saarland University, Germany
{mmosbach,babdullah,dklakow}@lsv.uni-saarland.de
s.degaetano@mx.uni-saarland.de
mariepauline.krielke@uni-saarland.de

## Abstract

Transformer-based language models achieve high performance on various tasks, but we still lack understanding of the kind of linguistic knowledge they learn and rely on. We evaluate three models (BERT, RoBERTa, and ALBERT), testing their grammatical and semantic knowledge by sentence-level probing, diagnostic cases, and masked prediction tasks. We focus on relative clauses (in American English) as a complex phenomenon needing contextual information and antecedent identification to be resolved. Based on a naturalistic dataset, probing shows that all three models indeed capture linguistic knowledge about grammaticality, achieving high performance. Evaluation on diagnostic cases and masked prediction tasks considering fine-grained linguistic knowledge, however, shows pronounced *model-specific* weaknesses especially on semantic knowledge, strongly impacting models' performance. Our results highlight the importance of (a) model comparison in evaluation task and (b) building up claims of model performance and the linguistic knowledge they capture beyond purely probing-based evaluations.

## 1 Introduction

Endeavors to better understand transformer-based masked language models (MLMs), such as BERT, are ever growing since their introduction in 2017 (cf. Rogers et al. (2020) for an overview). While the BERTology movement has enhanced our knowledge on the reasons behind BERT's performance in various ways, still plenty remains unanswered. Less well studied and challenging are linguistic phenomena, where, besides contextual information, identification of an antecedent is needed, such as relative clauses (RCs). Kim et al. (2019), e.g., analyzed BERT's comprehension of function words, showing how relativizers and prepositions are quite challenging for BERT. Similarly, Warstadt and Bowman (2019) find RCs to be difficult for BERT in the CoLA acceptability tasks. In this paper, we focus on RCs in American English to further enhance our understanding of the grammatical and semantic knowledge captured by pre-trained MLMs, evaluating three models: BERT, RoBERTa, and ALBERT. For our analysis, we train probing classifiers, consider each models' performance on diagnostic cases, and test predictions in a masked language modeling task on selected semantic and grammatical constraints of RCs.

RCs are clausal post-modifiers specifying a preceding noun phrase (antecedent) and are introduced by a relativizer (e.g., *which*). Extensive corpus research (Biber et al., 1999) found that the overall most common relativizers are *that*, *which*, and *who*. The relativizer occupies the subject or object position in a sentence (see examples (1-a) and (1-b)). In subject RCs, the relativizer is obligatory (Huddleston and Pullum, 2002, 1055), while in object position omission is licensed (e.g., *zero* in example (1-b)).

(1)    a.    *Children who eat vegetables are likely to be healthy. (subject relativizer, relativizer is obligatory)*

          b.    *This is the dress [that/which/zero] I brought yesterday. (object relativizer, omission possible)*

---

\* equal contribution.

Relativizer choice depends on an interplay of different factors.[1]  Among these factors, the animacy constraint (Quirk, 1957) is near-categorical: for animate head nouns the relativizer *who* (see Example 1) is strongly prioritized (especially over *which*) (D'Arcy and Tagliamonte, 2010).

Our aims are (1) to better understand whether sentence representations of pre-trained MLMs capture grammaticality in the context of RCs, (2) test the generalization abilities and weaknesses of probing classifiers with complex diagnostic cases, and (3) test prediction of antecedents and relativizers in a masked task considering also linguistic constraints. From a linguistic perspective, we ask whether MLMs correctly predict (a) grammatically plausible relativizers given certain types of antecedents (animate, inanimate) and vice versa grammatically plausible antecedents given certain relativizers (*who* vs. *which/that*), and (2) semantically plausible antecedents given certain relativizers considering the degree of specificity of predicted antecedents in comparison to target antecedents (e.g. *boys* as a more specific option than *children* in Example (1)). Moreover, we are interested in how these findings agree with probing results and investigate model specific behavior, evaluating and comparing the recent pre-trained MLMs: BERT, RoBERTa, and ALBERT. This is to our knowledge the first attempt comparing and analyzing performance of different transformer-based MLMs in such detail, investigating grammatical and semantic knowledge beyond probing.

Our main contributions are the following: (1) the creation of a naturalistic dataset for probing, (2) a detailed model comparison of three recent pre-trained MLMs, and (3) fine-grained linguistic analysis on grammatical and semantic knowledge. Overall, we find that all three MLMs show good performance on the probing task. Further evaluation, however, reveals model-specific issues with wrong agreement (where RoBERTa is strongest) and distance between antecedent-relativizer and relativizer-RC verb (on which BERT and ALBERT are better). Considering linguistic knowledge, all models perform better on grammatical rather than semantic knowledge. Out of the relativizers, *which* is hardest to predict. Considering model-specific differences, BERT outperforms the others in predicting the actual targets, while RoBERTa captures best grammatical and semantic knowledge. ALBERT performs worst overall.

## 2 Background

### 2.1 Models

BERT (Devlin et al., 2019) is a transformer-based (Vaswani et al., 2017) bidirectional network trained on masked language modeling and next-sentence-prediction. The extent to which BERT captures linguistic knowledge is widely studied in previous works (see §2.2). RoBERTa (Liu et al., 2019b) differs from BERT in four important aspects: trained on more data, no next-sentence-prediction objective, achieves lower perplexity on the training data, while using larger vocabulary. Given RoBERTa's superior performance over BERT on the GLUE benchmark (Wang et al., 2018), RoBERTa has replaced BERT as the model of choice in several recent studies investigating MLMs (Pruksachatkun et al., 2020; Talmor et al., 2020). Nevertheless, RoBERTa's linguistic properties and how they differ from BERT's remain relatively unexplored. ALBERT (Lan et al., 2020) is a recently proposed alternative to BERT and RoBERTa. It uses weight-sharing across all hidden layers — effectively applying the same non-linear transformation at every layer — and factorizes the embedding matrix into two separate matrices, resulting in significantly fewer model parameters compared to BERT and RoBERTa. ALBERT was shown to outperform BERT and RoBERTa when fine-tuned on several English NLP downstream tasks (Lan et al., 2020). However, to our knowledge, no previous work has systematically evaluated the linguistic knowledge of ALBERT. For all models, we consider the base variants: *BERT-base-cased*, *RoBERTa-base*, *ALBERT-base-v1* with 110M, 125M, and 12M parameters, respectively.

---

[1]These factors include register (fiction, news, academic texts), restrictiveness (restrictive RCs add information about the head noun necessary for identification of the latter (see Example (1)); non-restrictive RCs add information elaborating on a head noun which "is assumed to be already known" Biber et al. (1999, 602) and are usually separated by a comma, e.g. *My children, who I love, eat vegetables.*), animacy of the head noun (animate or inanimate antecedent), American (AE) vs. British English (BE), and definiteness of a pronominal antecedent (demonstrative vs. indefinite pronoun).

| Animate | Restrictive | SubjRC | Modification | Sentence |
|---|---|---|---|---|
| 1 | 1 | 1 | no modification | Katrina Haus was a woman **who** sought to attract stares, not turn them away. |
| 1 | 1 | 1 | who → which | *Katrina Haus was a woman **which** sought to attract stares, not turn them away. |
| 0 | 1 | 0 | no modification | She pulls out a course catalog, various forms, and a letter **which** she hands to Kevin. |
| 0 | 1 | 0 | which → that | She pulls out a course catalog, various forms, and a letter **that** she hands to Kevin. |
| 0 | 1 | 0 | no modification | Never permit your muzzle to cover anything **which** you are unwilling to shoot. |
| 0 | 1 | 0 | relativizer omission | Never permit your muzzle to cover anything you are unwilling to shoot. |
| 0 | 0 | 0 | no modification | I never saw a penny in royalties, **which** was all right with me. |
| 0 | 0 | 0 | which → who | *I never saw a penny in royalties, **who** was all right with me. |

Table 1: Examples from the dataset (minimal pairs). The relativizer is shown in bold. Modifying a sentence does not always result in an ungrammatical sentence. For example, when Restrictive=1 and SubjRC=0, relativizer omission yields a grammatical sentence.

## 2.2 Related Work

Related work has investigated grammaticality of unidirectional language models using *Minimal Pair Evaluation* (Marvin and Linzen, 2018; Wilcox et al., 2019; Warstadt et al., 2020; Hu et al., 2020a,b). Out of work on MLM prediction-based evaluation (e.g., Goldberg (2019); Ettinger (2020); Petroni et al. (2019); Jiang et al. (2020); Kassner and Schütze (2020)), only Goldberg (2019) has so far evaluated MLM predictions in the context of grammaticality. Our study adds to this line of work, focusing on RCs combining MLM prediction based evaluation with probing.

Work related to evaluating sentence embeddings considered prediction of sentence length, word content, and word order (Adi et al., 2016). Conneau et al. (2018) investigated an even broader range of linguistic properties. Extensive work has been done on probing pre-trained MLMs, especially BERT, for syntactic and semantic knowledge (see e.g., Tenney et al. (2019); Liu et al. (2019a) and Rogers et al. (2020) for a more comprehensive overview).

Most similar to our work is Warstadt et al. (2019), focusing on comparing evaluation methods, including probing and MLM evaluation to assess how models encode linguistic features. We contribute to this strand of research by building datasets from naturalistic (rather than artificial as in Warstadt et al. (2019)) data and comparing three transformer models: BERT, RoBERTa, and ALBERT. Our focus is on RCs as challenging sentence types for pre-trained language models as shown by, e.g., Kim et al. (2019) (showing relativizers to be quite challenging for BERT) and Warstadt and Bowman (2019) (finding RCs difficult for BERT in the COLA probing tasks).

## 3 Probing MLMs Representations for Grammatical Knowledge of RCs

We train supervised *probing* classifiers (here: acceptability classifiers) to assess the linguistic knowledge contained in a model's representations, focusing on grammaticality. A model's grammaticality awareness should be reflected in the hidden representations produced by the model for that particular sentence. Hence, by training a classifier on top of these representations, we should be able to discriminate between grammatical and ungrammatical sentences based on their sentence embeddings.[2] Moreover, besides capturing knowledge on grammaticality, we also examine whether this knowledge becomes more or less separable in the representation, considering representations produced by different layers of a model.

### 3.1 Dataset Construction

To probe pre-trained MLMs' performance on sentences containing RCs, we construct a controlled set of 48,060 sentences and their acceptability labels using an automated procedure. Our probing dataset is a subset of naturally occurring sentences extracted from the fiction portion of the COCA corpus (Davies, 2015). First, we extract all sentences containing only one pronoun from the set {*who, whom, whose, which, that*}. We then parse the sentences using SpaCy dependency parser and keep only those sentences where the pronoun constitutes a relativizer (identified by the tag RELCL). From the parse tree, we also

---

[2]Here, we assume that grammatical and ungrammatical sentences are *linearly separable* in the embedding space.

determine whether the relativizer fills the subject or the object position in the RC. The automatic selection procedure is illustrated in Fig. 2 in §A.1.

On top of grammatical sentences from the corpus, we manipulate the data obtaining a set with unacceptable ones. To this end, we populate three Boolean meta-data variables for each grammatical sample in our data using a set of hand-crafted linguistic rules: ANIMATE, RESTRICTIVE, and SUBJRC. Based on the values of these three meta-data variables, a set of modifications is applied to convert a grammatical sentence into an ungrammatical one. The procedure of creating the dataset is explained in detail in §A.1. Our final dataset consists of 42.7k and 5.3k samples for training and evaluation, respectively. Both splits are balanced, i.e. accuracy of the majority baseline is 50%. Table 1 presents a set of minimal pair examples that are generated using our procedure.

## 3.2 Experimental Setup

We train logistic regression acceptability classifiers on sentence embeddings obtained from the hidden layers of a pre-trained model. We compute sentence embeddings of BERT, RoBERTa, and ALBERT and two non-contextualized baselines: GloVe embeddings (Pennington et al., 2014) trained on English Wikipedia and the Gigaword corpus and fasttext embeddings (Bojanowski et al., 2016) trained on English Wikipedia and News data (Mikolov et al., 2018). As an additional baseline we use a rule based classifier that simply classifies sentences containing a relativizer (who, which, that) as grammatical and as ungrammatical otherwise.

Input sentences are pre-processed by adding two special tokens, [CLS] and [SEP], at the beginning and end of each input sentence, respectively.[3] To construct sentence embeddings, we apply two different pooling strategies: CLS- and mean-pooling.[4] We obtain sentence embeddings from all hidden layers of the MLM models, including the non-contextualized embedding layer (layer 0). We treat accuracy on the acceptability classification task as a proxy for the linguistic knowledge encoded in a model's sentence embeddings, which we evaluate on a held-out test set. We use the huggingface transformers (Wolf et al., 2019) and flair (Akbik et al., 2018) libraries as well as Scikit-learn (Pedregosa et al., 2011) for obtaining embeddings and training the logistic regression classifiers. Code to reproduce our probing dataset and results is available online: https://github.com/uds-lsv/rc-probing.
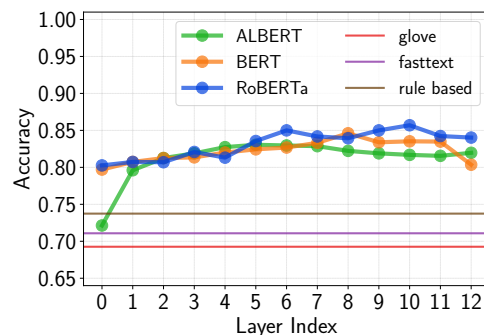


Figure 1: Layer-wise probing accuracy on the test set for pre-trained transformer and baseline models using mean-pooling.

## 3.3 Probing Results and Discussion

Fig. 1 shows the layer-wise probing accuracy for all models when using mean-pooling. A side-by-side comparison with CLS-pooling is shown in Fig. 4 in §A.2. Mean-pooling leads to significantly higher probing accuracies for all models, suggesting a sub-optimal encoding of sentence-level information in the CLS token representation. Hence, we stick to mean-pooling in the following. We find that the rule base classifier is a surprisingly strong baseline, outperforming both the GloVe and fasttext baselines. Moreover, Fig. 1 shows that all three transformer-based models improve significantly over the baselines for almost all layers. The only exception is layer 0 of ALBERT, which performs similarly to the fasttext baseline and worse than the rule based classifier. We attribute this finding to the embedding factorization of ALBERT. At lower (contextualized) layers (1–5), probing accuracies of BERT, RoBERTa, and ALBERT are almost identical. For higher layers, both BERT and RoBERTa improve over ALBERT, with ALBERT's probing accuracy remaining roughly constant. We note that ALBERT has significantly less parameters than BERT and RoBERTa (12M vs. 110M and 125M), which might explain the lower

---

[3]To be precise, for RoBERTa the special tokens are <s> and </s>. For GloVe and fasttext embeddings as well as the rule base classifier, we tokenize on the word level without using special tokens.

[4]CLS-pooling simply returns the vector representation of the first token of the input sentence, i.e. the [CLS] token. Mean-pooling computes a sentence embedding by taking the mean over all (sub-word) token representations of the input sentence.

| Modification | GloVE | fasttext | rule based | BERT | RoBERTa | ALBERT |
|---|---|---|---|---|---|---|
| no modification | 67.3 | 69.2 | 100 | 83.7 (78.4) | 85.3 (79.7) | 83.6 (71.5) |
| relativizer omission | 77.2 | 85.3 | 97.6 | 96.0 (98.3) | 95.1 (98.2) | 96.4 (94.8) |
| who → which | 88.0 | 81.7 | 0.00 | 84.9 (88.4) | 87.3 (89.6) | 78.8 (68.3) |
| which → who | 18.2 | 21.3 | 0.00 | 47.9 (1.73) | 50.0 (0.02) | 44.2 (18.7) |
| which → that | 12.8 | 8.0 | 100 | 80.0 (52.0) | 77.6 (44.8) | 72.0 (22.4) |
| total | 69.2 | 71.0 | 73.7 | 84.6 (79.7) | 85.5 (80.1) | 83.0 (72.1) |

Table 2: Test accuracy (in %) grouped by modification type (cf. Table 8 for statistics). For BERT, RoBERTa, and ALBERT we select the best model according to the probing results shown in Fig. 1. Numbers in parenthesis show the accuracy of the non-contextualized baseline (layer 0) for each model.

probing accuracy. We provide a more detailed discussion, investigating the role of the number of parameters, in §A.2.1. Overall, the fact that probing accuracy is above 80% for almost all layers suggests a reasonable encoding of sentence-level linguistic knowledge relevant for grammaticality classification in all pre-trained models. Notably, linear separability with respect to grammaticality emerges very early in the sentence embeddings of all three models.

To get a better understanding of the accuracy achieved by each of the models, we select the best classifier according to the results in Fig. 1 and report test accuracy grouped by modification in Table 2. For comparison, we additionally report accuracies of the non-contextualized baselines (layer 0) of each model in parentheses. The results show that while contextualization leads to higher probing accuracy overall, it is especially important for the which → who and which → that samples. From a linguistic viewpoint this is surprising: e.g. replacing *which* by *who* clearly makes a sentence ungrammatical and is typically easy to detect for humans. When looking at the training data, however, this observation is not surprising at all. The vast majority of sentences (15k samples) belong to the no modification group contain *who* as a relativizer. All of these sentences are grammatical (cf. Table 8). On the other hand, which → who contains only 2.5k samples, all of them are ungrammatical. Hence, our results in Table 2 reveal that the non-contextualized baselines might learn a simple heuristic, classifying all sentences that contain *who* as a relativizer as grammatical. The results of the rule based baseline classifier give further evidence for this interpretation, showing that a simple classifier which bases its predictions only on the existence of a relativizer has a surprisingly high accuracy (comparable to the non-conextualized baselines) on our dataset. Interestingly, BERT and RoBERTa seem to be especially susceptible to this short-cut learning as shown by the results of the non-contextual baselines for BERT and RoBERTa on the relativizer omission and which → who modifications.

## 3.4 Diagnostics

To investigate generalization abilities and model specific behavior of the best probing classifiers, we evaluate them on a diagnostics dataset containing sentences with the following properties: (1) adjacent antecedent and relativizer (see example (2-a), grammatical), (2) longer distance between antecedent and relativizer (see example (2-b), grammatical), (3) longer distance between relativizer and RC verb (see example (2-c), grammatical), (4) wrong agreement between adjacent antecedent and relativizer (see example (2-d), ungrammatical), and (5) intervening agreement attractors leading to wrong agreement (see (2-e), ungrammatical). To create the dataset, we manually select four sentences and manipulate these according to each of the above mentioned cases. Three sentences have nominal antecedents and one has a clausal antecedent. For the grammatical manipulations (case 1–3), we additionally test between restrictive/non-restrictive variants of each sentence. Overall, we test a total of 32 sentences[5] and evaluate the models' confidence based on the un-normalized log probabilities (logit), where predictions $> 0$ result in classifying sentences as grammatical. In the manipulations according to case 2, besides considering nominal vs. clausal antecedent, we look at the number of words of intervening phrases (length of 3–7 intervening words).[6]

---

[5] We opted for a small set for high control over factors influencing a choice.

[6] Note that for case 3, distance between relativizer and verb is always three words, so length is not a factor to be evaluated.

| Case | Factor | BERT | | ALBERT | | RoBERTa | |
|---|---|---|---|---|---|---|---|
| | | restrictive | non-restrictive | restrictive | non-restrictive | restrictive | non-restrictive |
| 1 (+) | nominal | 2.74 | 1.62 | 2.35 | 1.75 | −0.79 | −0.85 |
| | clausal | −0.39 | 0.89 | −1.18 | 0.57 | −0.85 | −0.57 |
| 2 (+) | nominal | 2.34 | 1.74 | 1.30 | 1.72 | −1.03 | −0.87 |
| | clausal | −0.39 | 0.89 | −1.18 | 0.57 | −0.85 | −0.57 |
| | 3–4 words | 2.35 | 2.52 | 1.30 | 2.55 | −1.03 | −0.55 |
| | > 4 words | −0.69 | 0.54 | −1.15 | 0.31 | −1.42 | −1.04 |
| 3 (+) | nominal | 1.99 | 1.97 | 2.04 | 1.78 | −1.06 | −0.80 |
| | clausal | −0.84 | 0.43 | −1.54 | 0.17 | −0.89 | −0.35 |
| 4 (*) | nominal | −0.47 | – | −0.40 | – | −1.46 | – |
| | clausal | −1.38 | – | −1.75 | – | −1.24 | – |
| 5 (*) | nominal | 0.56 | – | 0.08 | – | −1.24 | – |
| | clausal | −1.38 | – | −1.75 | – | −1.24 | – |

Table 3: Prediction confidence in mean logit ($> 0$: grammatical, $< 0$: ungrammatical). Sentences in case 1 to 3 should be predicted as grammatical (+) and case 4 and 5 as ungrammatical (*).

(2)  a.  *We just heard **a debate which** was about the differences in wage rates [...]. (case 1: adjacent antecedent and relativizer)*

b.  *[...] **a debate** on one of the most famous television channels **which** was about [...]. (case 2: distance antecedent to relativizer)*

c.  *[...] a debate **which** in many regards **was** an important one about [...]. (case 3: distance relativizer to verb)*

d.  *[...] **a debate who** was about [...]. (case 4: wrong agreement antecedent-relativizer)*

e.  *[...] a debate **by DeGeneres** who was about [...]. (case 5: agreement attractor)*

**Case 1.** For restrictive and non-restrictive variants and nominal antecedents, BERT and ALBERT correctly predict with relatively high confidence the sentences to be grammatical (restrictive: 2.7, non-restrictive: 1.6). For the clausal antecedent, they fail in the restrictive variant (−0.39 and −1.18, respectively), while they can deal with non-restrictive RCs (0.89 and 0.57), the comma possibly being a strong indicator of non-restrictiveness. RoBERTa always wrongly predicts ungrammaticality.

**Case 2.** With intervening phrases between antecedent and relativizer (see example (2-b)), BERT and ALBERT again deal well in both restrictive and non-restrictive RCs for nominal antecedents, but fail for restrictive RCs in the clausal case. RoBERTa again fails to predict the grammatical class. Considering the length of the intervening phrase, for restrictive RCs, BERT deals with intervening phrases with high confidence (2.35), provided they are relatively short (3–4 words). For longer distance, BERT is less confident (−0.69) and chooses the (wrong) ungrammatical class. ALBERT acts similarly, but fails with higher confidence (−1.15). RoBERTa always fails without a clear pattern. For the non-restrictive RCs, both BERT and ALBERT correctly predict the grammatical class, even though distance affects their confidence (the longer the less confident: 2.52 for shorter, 0.54 for longer phrases). RoBERTa again predicts the ungrammatical class for all sentences, but is especially confident for greater distance between antecedent and relativizer (−0.55 for shorter, −1.04 for longer phrases).

**Case 3.** With restrictive RCs, RoBERTa has problems with intervening phrases between relativizer and RC verb, again always predicting the ungrammatical class. BERT predicts this case correctly, but fails at sentences with clausal antecedent, even though with lower confidence than ALBERT (−0.84 vs. −1.54). In non-restrictive RCs, BERT and ALBERT obtain perfect accuracy for both nominal and clausal antecedents, being quite confident in their predictions, while RoBERTa still fails with sometimes very high confidence.[7]

**Case 4.** With wrong agreement and adjacent nominal antecedent and relativizer, RoBERTa is quite confident of the ungrammaticality of the sentences (−1.46), while BERT and ALBERT are confident in

---

[7]BERT and ALBERT seem to rely on the comma as an indicator of grammaticality for non-restrictive RCs (cases 0–2).

| | | objRC | | | subjRC | | |
|---|---|---|---|---|---|---|---|
| | | who | which | that | who | which | that |
| **MP@1** | BERT | **0.80** | 0.29 | 0.89 | **0.98** | 0.02 | **0.92** |
| | RoBERTa | **0.80** | 0.24 | **0.96** | 0.92 | 0.04 | **0.92** |
| | ALBERT | 0.32 | **0.43** | 0.86 | 0.96 | **0.16** | 0.76 |
| **MTR** | BERT | **1.23** | 1.83 | 1.11 | **1.00** | 2.46 | **1.10** |
| | RoBERTa | **1.23** | 1.95 | **1.04** | 1.08 | 2.18 | 1.22 |
| | ALBERT | 2.29 | **1.73** | 1.39 | 1.02 | **1.18** | 1.37 |
| **MNE** | BERT | 0.09 | 0.08 | **0.05** | 0.05 | 0.07 | 0.06 |
| | RoBERTa | **0.08** | **0.06** | 0.05 | 0.04 | 0.05 | 0.04 |
| | ALBERT | 0.19 | 0.21 | 0.18 | 0.12 | 0.13 | 0.17 |

(a) Metrical evaluation (MP@1: mean precision at 1, MTR: mean target rank, MNE: mean normalized entropy)

| | | objRC | | | subjRC | | |
|---|---|---|---|---|---|---|---|
| | | who | which | that | who | which | that |
| **AN** | BERT | 0.87 | **1.00** | 0.93 | **1.00** | 0.94 | 0.96 |
| | RoBERTa | 0.81 | **1.00** | 0.96 | 0.94 | 0.90 | 0.94 |
| | ALBERT | 0.71 | **1.00** | 0.93 | **1.00** | 0.94 | 0.92 |
| **PL** | BERT | 0.81 | 0.98 | 0.93 | **1.00** | 0.92 | 0.96 |
| | RoBERTa | **0.97** | **0.98** | 0.96 | 0.98 | **0.98** | **0.98** |
| | ALBERT | 0.68 | 0.95 | 0.89 | 0.98 | 0.96 | 0.92 |
| **GR** | BERT | 0.94 | **1.00** | 0.93 | **1.00** | 0.98 | **1.00** |
| | RoBERTa | 0.97 | **1.00** | 0.96 | **1.00** | 0.98 | **1.00** |
| | ALBERT | 0.68 | 0.95 | 0.89 | 0.98 | **0.98** | 0.96 |

(b) Semantic (AN: animacy, PL: plausibility) and grammaticality (GR) evaluation.

Table 4: *Relativizer* prediction: Quantitative (a) and qualitative (b) evaluation. **Bold**: best result for each metric and category across models. Underline: best result for each model per metric across categories.

the clausal antecedent case ($-1.38$ and $-1.75$, respectively), but much less confident or even wrong with nominal antecedents.

**Case 5.** With intervening agreement attractors and wrong agreement, BERT gets confused due to the attractors (see example (2-e), where *DeGeneres* is considered the antecedent instead of *debate*) and is only confident in the clausal antecedent case ($-1.38$). RoBERTa, instead, is very confident in recognizing the wrong agreement ($-1.24$), while ALBERT gets confused as well, even though less and with much less confidence than BERT ($0.08$).

In summary, RoBERTa is quite confident in wrong agreement cases. BERT and ALBERT deal much better than RoBERTa with longer distances between antecedent and relativizer. Also, BERT and ALBERT learn to recognize non-restrictive RCs quite well and can deal with phrases between relativizer and RC verb. Thus, even though the models achieve very high probing accuracy overall (see § 3.3), evaluating on more complex cases reveals that each model seems to rely on different kinds of information, strongly affecting the generalization abilities of the probing classifiers. While we are aware of the diagnostics set's very limited extend, hindering generalizable conclusions, the controlled diagnostics evaluation gives an indication of possible differences underlying prediction choices across models.

## 4 Analyzing MLMs Predictions for RC Awareness

### 4.1 Motivation

For a more comprehensive picture of the differences between models, we perform a masked language modeling evaluation (Goldberg, 2019), looking at the models' predictions of relativizers as well as antecedents. Besides grammaticality, we also test whether the models capture semantic knowledge.

### 4.2 Analyzing Grammatical and Semantic Knowledge

We extract sentences containing restrictive[8] object or subject RCs with one of the three relativizers from the registers magazines and academic[9] from the COCA corpus (Davies, 2015). All sentences are formatted as described in § 3.2, e.g. [CLS] The woman [MASK] studies linguistics. [SEP]. The size of each set of RC type depends on the frequency of available sentences in the COCA corpus and therefore varies between types from 20 to 50 sentences.

### 4.2.1 Relativizer Prediction

We test all three models by masking the relativizer, considering the following metrics for evaluation: (1) mean precision at 1 (MP@1) (model's precision of target prediction at first position), (2) mean target rank (MTR), and (3) normalized mean entropy (NME) (uncertainty of the model's prediction). For

---

[8]We deliberately exclude non-restrictive RCs due to the indicative comma facilitating RCs recognition (see § 3.4).

[9]We exclude fiction as it was used for the probing experiments, and spoken, due to possible high noise in spoken data.

MP@1, all three models generally perform well across RCs apart from *which* (Table 4a). BERT and RoBERTa show similar performance, while ALBERT diverges slightly, with fairly weak predictions of *who* in object RCs and comparatively accurate predictions of *which* in both object and subject RCs. MTR is negatively correlated with MP@1 (the larger the divergence from 1, the lower the precision of the prediction). NME reflects the two other measures: BERT and RoBERTa are quite confident about their predominantly successful predictions, while ALBERT shows a higher uncertainty about overall much weaker predictions.

Next, we manually evaluate[10] the actual predictions according to three criteria: *animacy* (agreement between antecedent and relativizer), *plausibility* (semantically plausible sentence) and *grammaticality* (Table 4b). Results show that the actual felicity of the predictions is much stronger than MP@1 would suggest, since non-target predictions are not necessarily infelicitous. This is especially true for *which*, due to high interchangeability with *that*. While both relativizers are synonymous, *which* is primarily used in non-restrictive RCs. Since the sample sentences are restrictive, the models seem to predict the most frequent relativizer in restrictive RCs, which is *that* (see example (3)).

(3)    *The action [MASK] it contemplates is command. (all models) (object RC, target= which, prediction= that)*

Animacy is predicted reliably by BERT and RoBERTa, while ALBERT seems to have issues with *who* object predictions. This is due to ALBERT's preference to predict *that* instead of *who*, especially in fuzzy cases of general nouns describing humans (*person, people, family, etc.*). Sometimes animacy is predicted correctly, however, case marking is infelicitous; the model seems to only take into account the subject and verb of the clause ignoring possible clausal complements (see example (5)).

(4)    *He spared no expense in moving, reassembling, and restoring buildings of people [MASK] he felt were the backbone of our nation. (ALBERT) (object RC, target= who, prediction=that)*

(5)    *Jennifer had been having an online affair with a person [MASK] she believed was a man named Christopher. (ALBERT) (object RC, target= who, prediction= whom)*

Accuracy for plausibility is very high, too, with ALBERT making the least plausible predictions. Note that plausibility entails grammaticality. Example (5) has animacy agreement, but is neither grammatical nor plausible. Grammaticality in contrast does not entail plausibility (see example (6)). Here, BERT predicts the most frequent collocate of *whirl* according to COCA (Davies, 2015).

(6)    *This is something else, for I saw a whirl [MASK] I knew was a large bass. (BERT) (object RC, target= which, prediction= wind)*

Of all criteria, predictions are most accurate for grammaticality. Even ALBERT reaches a precision of 90% on average. Finally, we consider the ratio of (any) relativizers predicted vs. other word class, indicating how well the models recognize a syntagmatic environment typical for a RC to occur. RoBERTa, again, is most successful here (95% of predictions are relativizers), followed by BERT (94%) and ALBERT (90%). In summary, BERT and RoBERTa perform equally well at target prediction, while RoBERTa is most successful qualitatively (grammaticality, plausibility, and relativizer prediction). Overall, our qualitative analysis shows that all models largely make grammatical, plausible and animacy-conform predictions.

### 4.2.2 Antecedent Prediction

Here, we test all three models by masking the antecedent (see example (7)), considering again, mean precision (MP@1), mean target rank (MTR), and normalized mean entropy (NME) (Table 5a). Due to higher variation in lexical choices for antecedents, MP@1 is expectedly much lower than for relativizers. Best @1 predictions are achieved by all models for *who* and *that* object RCs. Comparing models, BERT achieves best performance in all cases. MTR is around 2.5 and 3.5, with antecedents in *who* subject

---

[10]Evaluations were done by two linguistic experts.

|  |  | objRC | | | subjRC | | |
|---|---|---|---|---|---|---|---|
|  |  | who | which | that | who | which | that |
| **MP@1** | BERT | **0.38** | **0.22** | **0.38** | 0.27 | **0.31** | 0.29 |
|  | RoBERTa | 0.31 | 0.08 | 0.29 | 0.22 | 0.17 | 0.14 |
|  | ALBERT | 0.28 | 0.14 | 0.29 | 0.20 | 0.15 | 0.14 |
| **MTR** | BERT | 3.04 | **2.82** | 3.41 | 2.42 | 3.30 | 2.93 |
|  | RoBERTa | **2.65** | 3.80 | 2.33 | 2.60 | **2.67** | **2.82** |
|  | ALBERT | 3.20 | 3.23 | 2.45 | 2.58 | 3.17 | 3.18 |
| **NME** | BERT | 0.23 | 0.38 | 0.27 | 0.26 | 0.32 | 0.36 |
|  | RoBERTa | **0.18** | **0.28** | 0.22 | 0.21 | 0.24 | 0.27 |
|  | ALBERT | 0.40 | 0.50 | 0.41 | 0.39 | 0.42 | 0.50 |

(a) Metrical evaluation (MP@1: mean precision at 1, MTR: mean target rank, MNE: mean normalized entropy)

|  |  | objRC | | | subjRC | | |
|---|---|---|---|---|---|---|---|
|  |  | who | which | that | who | which | that |
| **AN** | BERT | 0.97 | 0.94 | **0.95** | 0.98 | 0.94 | 0.98 |
|  | RoBERTa | **1.00** | 0.97 | **0.95** | **1.00** | **1.00** | 0.96 |
|  | ALBERT | **1.00** | **1.00** | 0.90 | 0.96 | **1.00** | 0.96 |
| **PL** | BERT | **0.97** | 0.94 | 0.95 | **1.00** | 0.90 | **0.96** |
|  | RoBERTa | 0.97 | 0.97 | **1.00** | 0.98 | **0.92** | 0.92 |
|  | ALBERT | 0.97 | 0.97 | 0.76 | 0.96 | 0.65 | 0.73 |
| **GR** | BERT | **1.00** | 0.97 | **1.00** | **1.00** | 0.98 | **1.00** |
|  | RoBERTa | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
|  | ALBERT | **1.00** | **1.00** | 0.95 | 0.96 | 0.98 | 0.94 |

(b) Semantic (AN: animacy, PL: plausibility) and grammaticality (GR) evaluation.

Table 5: *Antecedent* prediction: quantitative (a) and qualitative (b) evaluation. **Bold**: best result for each metric and category across models. <u>Underline</u>: best result for each model per metric across categories.

RCs being predicted most easily by all models (rank around 2.5), also reflected in NME. The models are equally confident in *who* object RCs, while they are less confident for the other cases, with ALBERT being the least confident and RoBERTa the most confident model.

(7) *Rheumatologists have to be medical detectives, because so many of the [MASK] **that** we treat are obscure. (object RC, masked target: **diseases**)*

Considering the first predicted antecedent, we qualitatively evaluate whether (a) the animacy constraint is met, i.e. *who* RCs should have animate antecedents, and *that* and *which* most prominently inanimate ones[11] (see (8)), (b) the sentence is plausible, (c) the sentence is grammatical (see example (9) being ungrammatical in noun-verb agreement of *animals* (plural) and *was* (singular)[12]).

(8) *They picked a great foreman in the middle-aged, an African-American [MASK] **who** she said is a science professor with a Ph.D. (target = man, prediction = comma, animate)*

(9) *By the time Ganivet drafted Idearium español, he was applying Ribot's psychological theories to the apparent inertia of [MASK] **which** he deemed was suffering from a collective psychological malady. (target = Spain, prediction = animals)*

Overall, BERT and RoBERTa show very high accuracy for these cases (Table 5b). In few cases, the animacy of the antecedent is changed (see example (10)) and sentences are grammatical but semantically implausible or ungrammatical and implausible (see example (10) and example (11)). ALBERT performs worst considering plausibility, especially for *which* in subject RCs (see example (12)), most cases being repetitions of words in the sentence (see example (13)).

(10) *The only thing that brightened the gloom stretching out before him was the goodness of the [MASK] **that** had offered him refuge before his trial. (target = family, prediction = darkness)*

(11) *Coffee, the cash, and the goods they purchased, even when used in traditional exchange, were devoid of the social relations with [MASK] **which** were present in traditional products. (target = predecessors, prediction = humans)*

(12) *He and Carolyn saw all of Kevin's college [MASK] **that** were within a day's drive of McIntyre. (target = games, prediction = campuses)*

(13) *The bill makes it illegal to adopt or enforce any law or [MASK] **which** allows gays to claim discrimination. (target = policy, prediction = law)*

---

[11]In cases where an elliptical version or other is predicted, we still check whether the animacy was kept or changed.

[12]Here, BERT seems to consider *inertia* as the antecedent, matching the agreement with *was*.

In summary, all models have problems with *which* relativizer, while they perform best with *who*. For animacy and plausibility the models suffer the most, especially for subject RCs, but perform quite well for grammaticality, indicating a better awareness for grammatical than semantic knowledge.

We further evaluate the models' semantic knowledge considering predicted types of antecedents ((a) identical to the target, (b) synonym, (c) hypernym, general noun, determiner, pronoun, (d) hyponym[13], or (e) not directly related (i.e. no direct hierarchical relationship) or completely unrelated[14]). Results (see Table 10, §A.3) show that all models perform less well on subject than object RCs. ALBERT performs worst as the majority falls into not directly related/unrelated antecedents or hypernyms (more general words). RoBERTa is quite good at predicting *identical* targets outperforming the other two models, especially in object RCs with *who*. In *who* subject RCs, BERT and ALBERT most often predict more general antecedents given more specific targets (hypernyms, e.g., *person* instead of *girl*, *workers*, etc.).

## 5 Discussion and Conclusion

With our work we moved towards tackling some issues in the evaluation of the linguistic capabilities of pre-trained transformer-based masked language models as, e.g., proposed by Rogers et al. (2020). Most prominently, we try to better understand how a strong performance on supervised probing tasks is reflected in the predictions of the language models. To do so, we create a dataset based on naturalistic (not artificially generated) data and perform an extensive evaluation of masked language predictions in the context of RCs. Moreover, rather than considering one model only, we compare three models to investigate the extend to which findings for BERT can be generalized to other transformer-based models.

Our probing results show a significant improvement of all three transformer-based models over the baselines for almost all layers, suggesting indeed encoding of linguistic knowledge relevant for grammaticality classification, as shown in previous work (e.g. Goldberg (2019)). Performance is similar across models, with BERT and RoBERTa performing slightly better than ALBERT, and while contextualization improves overall, we have shown it helps immensely when considering less frequent RC modification types such as `which` → `who`, for which uncontextualized baselines learn simple heuristics. Evaluation on a diagnostic set, however, clearly reveals weaknesses of probing classifiers and model-specific behavior. RoBERTa is quite confident in wrong agreement cases between antecedent and relativizer, while this is problematic for BERT and ALBERT. Considering distance between relativizer and antecedent/RC verb, they clearly outperform RoBERTa. Based on these insights, **we conclude that viewing probing results in isolation can lead to overestimating the linguistic capabilities of a model**.

Our masked language modeling evaluation provided deeper insights into model-specific differences. We evaluated relativizer as well as antecedent prediction. Overall, all models show better performance on grammatical than semantic knowledge (animacy and plausibility). Regarding relativizer prediction, all models perform worst on the target word *which* (plausible, as it is the most versatile of the relativizers). Comparing models, BERT is best in predicting the actual targets, RoBERTa outperforms the others in capturing grammatical and semantic knowledge, while ALBERT performs worst overall. Evaluation on semantic types of antecedents shows prediction of unrelated or not directly related antecedents, especially for ALBERT in *which* RCs. Interestingly, both BERT and ALBERT predict hierarchically more general antecedents in *who* RCs, while RoBERTa is able to better capture specificity.

We believe that more work in this direction will lead to (a) awareness of the complexity of linguistic knowledge that such models might have to capture, considering e.g. generalization tasks, and (b) provide enhancement in evaluation strategies on how to best capture linguistic knowledge, such as a combination of probing, diagnostic, and cloze tests, but also in developing best practices of evaluation.

## Acknowledgements

---

[13]Hyponyms are semantically lower in hierarchy and thus more specific.

[14]We relied on WordNet whenever possible (low coverage) and on other linguistic resources plus linguistic know-how of two linguists. The annotated data will be made available upon publication.

# References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*, 2016.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. *Longman Grammar of Spoken and Written English*. Longman, Harlow, UK, 1999.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*, 2016.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL https://www.aclweb.org/anthology/P18-1198.

Alexandra D'Arcy and Sali A. Tagliamonte. Prestige, accommodation, and the legacy of relative who. *Language in Society*, 39(3):383–410, 2010. doi: 10.1017/S0047404510000205.

Mark Davies. Corpus of Contemporary American English (COCA), 2015. URL https://doi.org/10.7910/DVN/AMUDUW.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Allyson Ettinger. 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, 2020. doi: 10.1162/tacl_a_00298. URL https://doi.org/10.1162/tacl_a_00298.

Yoav Goldberg. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*, 2019.

Jennifer Hu, Sherry Yong Chen, and Roger Levy. A closer look at the performance of neural language models on reflexive anaphor licensing. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 323–333, New York, New York, January 2020a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.scil-1.39.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 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, July 2020b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.158.

Rodney Huddleston and Geoffrey K. Pullum. *The Cambridge Grammar of the English Language*. Cambridge University Press, 2002. doi: 10.1017/9781316423530.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. doi: 10.1162/tacl_a_00324. URL https://www.aclweb.org/anthology/2020.tacl-1.28.

Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.698.

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. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-1026. URL `https://www.aclweb.org/anthology/S19-1026`.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=H1eA7AEtvS`.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1112. URL `https://www.aclweb.org/anthology/N19-1112`.

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

Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1151. URL `https://www.aclweb.org/anthology/D18-1151`.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL `https://www.aclweb.org/anthology/D14-1162`.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL `https://www.aclweb.org/anthology/D19-1250`.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online, July 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.acl-main.467`.

Randolph Quirk. Relative Clauses in Educated Spoken English. *English Studies*, 38:97–109, 1957.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works, 2020.

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. Teaching Pre-Trained Models to Systematically Reason Over Implicit Knowledge. *arXiv preprint arXiv:2006.06609*, 2020.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=SJzSgnRcKX`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf`.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL `https://www.aclweb.org/anthology/W18-5446`.

Alex Warstadt and Samuel R. Bowman. Linguistic analysis of pretrained sentence encoders with acceptability judgments, 2019.

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. 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 (EMNLP-IJCNLP)*, pages 2877–2887, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1286. URL `https://www.aclweb.org/anthology/D19-1286`.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020. doi: 10.1162/tacl_a_00321. URL `https://www.aclweb.org/anthology/2020.tacl-1.25`.

Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. Structural supervision improves learning of non-local grammatical dependencies. 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 3302–3312, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1334. URL `https://www.aclweb.org/anthology/N19-1334`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, pages arXiv–1910, 2019.

# A  Appendix

## A.1  Probing Dataset

Fig. 2 shows how to identify sentences containing RCs using a dependency parse tree, which were obtained using SpaCy. A sentence with an RC can be identified if the antecedent has an outgoing edge with the tag RELCL. The RC in the sentence can be further extracted since the main verb in the RC would have an incoming edge with the tag RELCL, and the subtree where RC main verb is the head constitutes the RC. This procedure enables us to extract sentences with RCs that are grammatical from text.
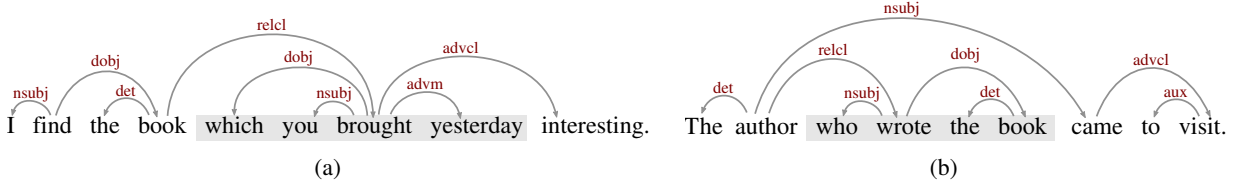


Figure 2: Visualized dependency parse trees extracted by SpaCy: (a) object RC, and (b) subject RC.

However, probing classifiers for grammatical acceptability require both grammatical and ungrammatical sentences. Since grammaticality of RCs in English depends on restrictiveness, animacy of the head noun, and whether the relativizer occupies the subject or the object position in the sentence, we populate three meta-data variables for each grammatical sentence with respect to these aspects of RCs: ANIMATE, RESTRICTIVE, and SUBJRC. The variables ANIMATE and RESTRICTIVE are populated using a set of hand-crafted rules for the usage of RCs in American English (illustrated in Fig. 3), while the variable SUBJRC is populated using the incoming edge to the relativizer in the parse tree (e.g. NSUBJ for subject RC vs. DOBJ for object RC). We discard all sentences where at least one meta-data variable cannot be populated with certainty.
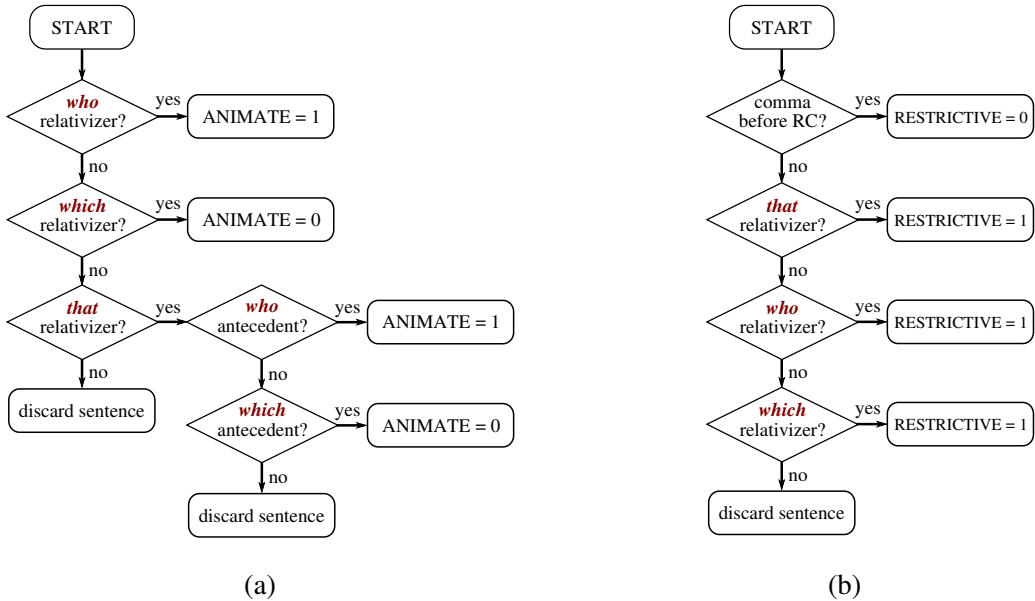


Figure 3: Annotation decision process for meta-data variables: (a) *Animate*, relativizer **who** and **which** are directly categorized as animate and non-animate, respectively. The relativizer **that** can be either way; thus, we categorize these sentences based on the antecedent. We compile two disjoints sets for antecedents that exclusively occur either with **who** or **which**, and the decision is made based on the membership of the antecedent. If the antecedent is not a member of either set, the sentence is discarded. (b) *Restrictive* can be easily identified since non-restrictive RCs in American English are always preceded by comma ",".

Using the aforementioned annotation procedure, and given the values of the three meta-data variables, each sentence can be manipulated to create an ungrammatical sentence that forms a minimal pair with the

| Animate | Restrictive | Subject RC | Modification | Label | Example |
|---|---|---|---|---|---|
| | | | None | 1 | Katrina Haus was a woman **who** sought to attract stares, not turn them away. |
| 1 | 1 | 1 | Relativizer omission | 0 | *Katrina Haus was a woman sought to attract stares, not turn them away. |
| | | | who → which | 0 | *Katrina Haus was a woman **which** sought to attract stares, not turn them away. |
| | | | None | 1 | Two people **who** I loved very much have left me . |
| 1 | 1 | 0 | Relativizer omission | 1 | Two people I loved very much have left me . |
| | | | who → which | 0 | *Two people **which** I loved very much have left me . |
| | | | None | 1 | Buck , **who** was snoozing over in the corner , woke up and looked around. |
| 1 | 0 | 1 | Relativizer omission | 0 | *Buck , was snoozing over in the corner , woke up and looked around. |
| | | | who →which | 0 | *Buck , **which** was snoozing over in the corner , woke up and looked around. |
| | | | None | 1 | Sally turned with a welcoming grin , expecting to see Gus , **whom** she liked a lot. |
| 1 | 0 | 0 | Relativizer omission | 0 | *Sally turned with a welcoming grin , expecting to see Gus , she liked a lot |
| | | | who →which | 0 | *Sally turned with a welcoming grin , expecting to see Gus , **which** she liked a lot |
| | | | None | 1 | One is a rather, um, disconcerting bit of information **which** has reached my ears. |
| 0 | 1 | 1 | Relativizer omission | 0 | *One is a rather, um, disconcerting bit of information has reached my ears. |
| | | | which → who | 0 | *One is a rather, um, disconcerting bit of information **who** has reached my ears. |
| | | | which → that | 1 | One is a rather, um, disconcerting bit of information **that** has reached my ears. |
| | | | None | 1 | Pulls out a course catalog, various forms, and a letter **which** she hands to Kevin. |
| 0 | 1 | 0 | Relativizer omission | 0 | *Pulls out a course catalog, various forms, and a letter she hands to Kevin. |
| | | | which → who | 0 | *Pulls out a course catalog, various forms, and a letter **who** she hands to Kevin. |
| | | | which → that | 1 | Pulls out a course catalog, various forms, and a letter **that** she hands to Kevin. |
| | | | None | 1 | I never saw a penny in royalties, **which** was all right with me. |
| 0 | 0 | 1 | Relativizer omission | 0 | *I never saw a penny in royalties, was all right with me. |
| | | | which → who | 0 | *I never saw a penny in royalties, **who** was all right with me. |
| | | | None | 1 | Lyric clips her Walkman to her fanny pack, **which** she wears pouch forward. |
| 0 | 0 | 0 | Relativizer omission | 0 | *Lyric clips her Walkman to her fanny pack, she wears pouch forward. |
| | | | which → who | 0 | *Lyric clips her Walkman to her fanny pack, **who** she wears pouch forward. |

Table 6: Examples of the generated paradigms and the possible modifications for each paradigm. It is worth pointing out that not all sentence modifications yield an unacceptable sentence. For example, the modification *relativizer omission* keeps the sentence grammatical in restrictive object RCs.

original sentence. The resulting paradigms from the three meta-data variables and the set of all possible modifications that be applied to each paradigm are presented in Table 6. Each grammatical sentence in the dataset is manipulated using all possible modifications of the paradigm to generate a "bag-of-sentences" associated with each sentence (where the original grammatical sentence is a member of). To create the final dataset, we sample one sentence from each bag and construct a balanced dataset of 48,060 sentences with grammatical acceptability labels. Tables 7 and 8 show summary statistics of the dataset.

| | Total | Acceptable | | Animate | | Restrictive | |
|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 0 | 1 | 0 | 1 |
| Training | 42720 | 21360 | 21360 | 11106 | 31614 | 25947 | 16773 |
| Test | 5340 | 2670 | 2670 | 1439 | 3901 | 3229 | 2111 |

Table 7: Summary statistics of the dataset splits used in this paper. The dataset is balanced with respect to acceptability judgment (the probing label **Acceptable**). **Animate** and **Restrictive** are meta-data variables, they are only used for creating the data and the modifications, not for probing.

| | No modif. | | Rel. omission | | who → which | | which → who | | which → that | |
|---|---|---|---|---|---|---|---|---|---|---|
| Acceptability | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Training | – | 20170 | 10512 | 116 | 8291 | – | 2557 | – | – | 1074 |
| Test | – | 2670 | 1299 | 19 | 1025 | – | 346 | – | – | 125 |

Table 8: Summary statistics of the dataset with respect to the modifications. Note that modifications are not balanced.

## A.2 Probing Results

Fig.4 shows layer-wise probing accuracies for all models considered in this paper. CLS-pooling clearly leads to worse sentence representations performing on par with the non-contextualized GloVe and fasttext mean-embeddings for most layers.
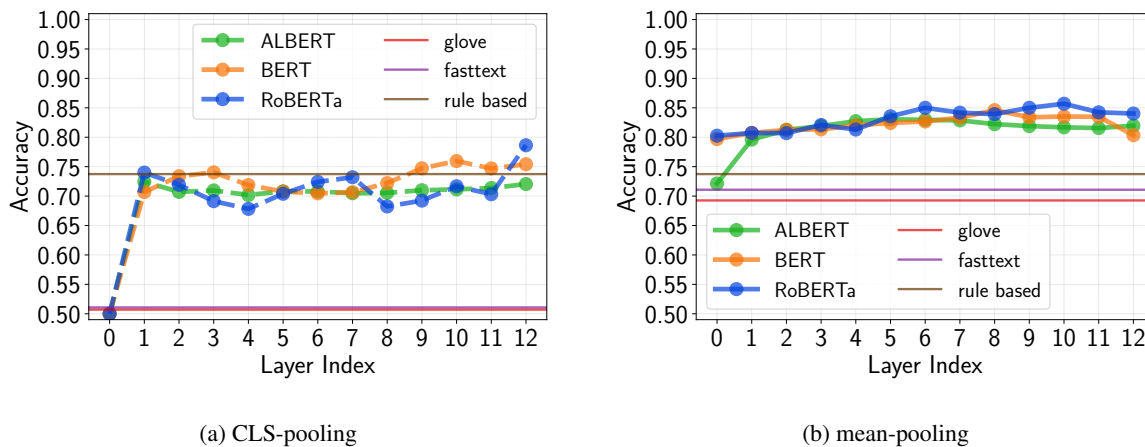


(a) CLS-pooling

(b) mean-pooling

Figure 4: Side-by-side comparison of layer-wise probing accuracy on the test set for pre-trained transformer and baseline models using (a) CLS-pooling and (b) mean-pooling.

### A.2.1 ALBERT-base-v1 vs. ALBERT-xxlarge-v1

Fig.5 shows layer-wise probing accuracies for ALBERT-base-v1 (ALBERT) and ALBERT-xxlarge-v1. ALBERT-xxlarge-v1 contains 235M parameters and hence roughly 10x as many as ALBERT-base-v1 and 2x as many as BERT and RoBERTa. As the results in Fig.5 show, the larger number of parameters indeed results in a significantly higher probing accuracy, outperforming all other models by a large margin. However, it should be noted that the larger number of parameters is the results of a larger dimensionality of ALBERT-xxlarge-v1's embeddings space, which is 5x larger that that of ALBERT-base-v1, BERT, and RoBERTa (4096 vs. 768). This makes the obtained probing results not directly comparable. We leave a more detailed investigation of the role of the size of the embedding space for future work.
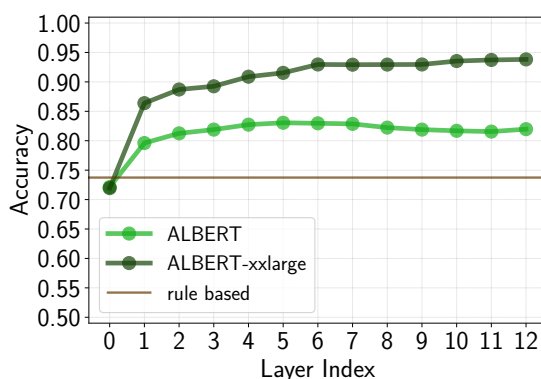


Figure 5: Comparison of ALBERT-base-v1 and ALBERT-xxlarge-v1 using mean-pooling. Table 9 shows the results of the best ALBERT-xxlarge-v1 probing classifier compared to all other models grouped by modification.

| Modification | GloVE | fasttext | rule based | BERT | RoBERTa | ALBERT | ALBERT-xxlarge |
|---|---|---|---|---|---|---|---|
| no modification | 67.3 | 69.2 | 100 | 83.7 (78.4) | 85.3 (79.7) | 83.6 (71.5) | 95.0 (70.6) |
| relativizer omission | 77.2 | 85.3 | 97.6 | 96.0 (98.3) | 95.1 (98.2) | 96.4 (94.8) | 96.2 (93.5) |
| who → which | 88.0 | 81.7 | 0.00 | 84.9 (88.4) | 87.3 (89.6) | 78.8 (68.3) | 94.6 (69.0) |
| which → who | 18.2 | 21.3 | 0.00 | 47.9 (1.73) | 50.0 (0.02) | 44.2 (18.7) | 80.6 (21.3) |
| which → that | 12.8 | 8.0 | 100 | 80.0 (52.0) | 77.6 (44.8) | 72.0 (22.4) | 73.6 (33.6) |
| total | 69.2 | 71.0 | 73.7 | 84.6 (79.7) | 85.5 (80.1) | 83.0 (72.1) | 93.8 (71.9) |

Table 9: Test accuracy (in %) grouped by modification type (cf. Table 8 for statistics). For BERT, RoBERTa, ALBERT-base-v1 (ALBERT) and ALBERT-xxlarge-v1 we select the best model according to the probing results shown in Fig. 1. Numbers in parenthesis show the accuracy of the non-contextualized baseline (layer 0) for each model. ALBERT-xxlarge-v1 performs especially well on the `which →` `who` modification.

## A.3 Qualitative Analysis for Predicted Type of Antecedent

Table 10 shows percentage for predicted types of antecedents: (a) identical to target, (b) synonym, (c) hypernym, general noun, determiner, pronoun, (d) hyponym (i.e. semantically lower in hierarchy and thus more specific), or (e) not directly related (i.e. no direct hierarchical relationship) or completely unrelated. Results show ALBERT to perform worst as the majority falls into not directly related/unrelated antecedents or hypernyms (more general words). RoBERTa is quite good at predicting *identical* targets outperforming the other two models, especially in object RCs with *who*. While the *which* case in object RCs is harder for all models, RoBERTa still makes best predictions considering identical and synonym prediction. All models perform less well on subject RCs, especially for *which* and *that* (40% to $> 50\%$ of unrelated/not directly related targets). Interestingly, for animate antecedents (*who* relativizer), the majority for all models falls into hypernyms, i.e. more general variants. This is especially the case for BERT and ALBERT.

| | type | objRC | | | subjRC | | |
|---|---|---|---|---|---|---|---|
| | | who | which | that | who | which | that |
| **BERT** | identical | **0.38** | 0.22 | **0.38** | 0.27 | 0.31 | 0.29 |
| | synonym | 0.03 | 0.08 | 0.10 | 0.10 | 0.06 | 0.02 |
| | hypernym | 0.07 | 0.19 | 0.14 | **0.40** | 0.17 | 0.16 |
| | hyponym | 0.17 | 0.00 | 0.10 | 0.15 | 0.06 | 0.08 |
| | unrelated/not directly related | 0.34 | **0.50** | 0.29 | 0.10 | **0.40** | **0.45** |
| **RoBERTa** | identical | **0.48** | 0.22 | **0.52** | 0.31 | 0.27 | 0.29 |
| | synonym | 0.00 | 0.11 | 0.00 | 0.10 | 0.15 | 0.08 |
| | hypernym | 0.21 | 0.19 | 0.05 | **0.31** | 0.08 | 0.08 |
| | hyponym | 0.17 | 0.00 | 0.10 | 0.12 | 0.08 | 0.08 |
| | unrelated/not directly related | 0.14 | **0.33** | 0.47 | 0.16 | **0.47** | **0.42** |
| **ALBERT** | identical | 0.28 | 0.14 | 0.29 | 0.20 | 0.15 | 0.14 |
| | synonym | 0.03 | 0.06 | 0.00 | 0.02 | 0.08 | 0.00 |
| | hypernym | 0.21 | 0.19 | 0.24 | **0.43** | 0.10 | 0.16 |
| | hyponym | 0.17 | 0.03 | 0.14 | 0.08 | 0.08 | 0.10 |
| | unrelated/not directly related | **0.31** | **0.58** | **0.33** | 0.27 | **0.58** | **0.59** |

Table 10: Predicted **types of antecedents** by relativizer in percentage. The type hypernym encompasses also general nouns, determiners, and pronouns.