Transformers in the loop: Polarity in neural models of language

Lisa Bylinina[*]
Bookarang, Amsterdam
bylinina@gmail.com

Alexey Tikhonov[*]
Yandex Technologies GmbH, Berlin
altsoph@gmail.com

Abstract

Representation of linguistic phenomena in computational language models is typically assessed against the predictions of existing linguistic theories of these phenomena. Using the notion of polarity as a case study, we show that this is not always the most adequate set-up. We probe polarity via so-called 'negative polarity items' (in particular, English *any*) in two pre-trained Transformer-based models (BERT and GPT-2). We show that – at least for polarity – metrics derived from language models are more consistent with data from psycholinguistic experiments than linguistic theory predictions. Establishing this allows us to more adequately evaluate the performance of language models and also to use language models to discover new insights into natural language grammar beyond existing linguistic theories. This work contributes to establishing closer ties between psycholinguistic experiments and experiments with language models.

1 Introduction

Recent Transformer-based language representation models (LRMs) – such as BERT and GPT-2 (Devlin et al., 2019; Radford et al., 2019) – show impressive results on practical text analysis tasks. But do these models have access to complex linguistic notions? The results in this domain are less clear – as well as ways to best approach this question.

Instead of asking whether LRMs encode fragments of current linguistic theory, we will directly compare metrics derived from LRMs to corresponding human judgments obtained in psycholinguistic experiments. The motivation for this is twofold. First, linguistic theories can be inaccurate – so, evaluating a model with respect to predictions of such theories is not informative about the model performance. Second, robust abstract theoretical notions rarely correspond to robust judgments in humans, and 'theoretical' and 'perceived' versions of the same phenomenon can be significantly different (for instance, see Geurts 2003 on inference judgments; discussed in Section 2). If this is something that LRMs inherit through training on human-produced texts, this makes LRMs an attractive possible component in an experimental pipeline, serving as a source of empirical predictions about human linguistic behaviour (Baroni, 2021; Linzen and Baroni, 2021).

As a case study, we focus on **polarity**: a complex property of sentences at the intersection of grammar and semantics. We tackle polarity via the distribution of items that are sensitive to it – namely, so-called **negative polarity items** (NPIs) like English *any*. As a basic illustration of NPI sensitivity to polarity, consider a pair of sentences in (1) (* = ungrammaticality):

(1) a. Mary didn't buy any books.
b. *Mary bought any books.

(1-a) is a negative sentence (has negative polarity), and *any* is grammatical in it. (1-b) is an affirmative sentence (has positive polarity) and *any* in this sentence is grammatically degraded compared to (1-a). Apart from this paradigmatic contrast, as we discuss below, polarity contrasts are expressed in a variety of ways and are tied to semantics.

As a proxy for a grammaticality measure, we will use the probability of `any` in the masked token position (in BERT) (following Goldberg 2019; Warstadt et al. 2019 a.o.) and perplexity increase when adding `any` to a sentence (in GPT-2). The differences in the metrics for the two different models stem from the differences in their architecture and training objectives. For all experiments, we use non-fine-tuned pre-trained LRMs. For this, we introduce our ANY dataset, which combines natural and synthetic data.

We find high levels of alignment between results of psycholinguistic experiments on monotonicity

[*]Equal contribution.

and NPIs, on the one hand – and our LRM-derived results, on the other hand. Furthermore, show how LRMs can be used to make new predictions about NPIs in contexts with different numerals and confirm these predictions in a psycholinguistic experiment.

This case study contributes to the complement of the 'interpretability of neural LRMs' research agenda: we can ask not only what linguistic tasks tell us about LRMs, but also what these models can help us find out about natural language (see Baroni 2021; Linzen and Baroni 2021 for a discussion along these lines).

The paper is structured as follows. First, in section 2, we set up the context for our study: we describe the background in theoretical and experimental linguistics in the domains relevant for our discussion. Section 3 discusses previous work on NPIs and polarity in computational linguistics. Section 4 contains the description of our experimental method. First, we introduce our ANY dataset; then, we describe the tests and metrics we use with BERT and with GPT-2 given our dataset. Section 5 discusses our results. In section 6, we go beyond state-of-the-art knowledge in experimental semantics and pragmatics and study the effect of the numeral on NPI acceptability – first, we do a BERT study and then confirm the results on human participants. Section 7 concludes: we propose directions for future work aligning experimental studies of language in humans and LRMs.

2 Background

NPIs are expressions with limited linguistic distribution. While their use is grammatical in some sentences, in other sentences their use results in ungrammaticality. The distribution of NPIs like *any* is governed by the notion of polarity that is much more intricate than the simple presence or absence of sentential negation, as in (1).

For instance, in examples (2)-(3), (2) are 'negative enough' to allow for (='license') *any*, while (3) are not – even though none of these sentences contain overt sentential negation.

(2) a. None of the boxes contain anything.
b. Nobody talked to anybody.
c. At most five students did anything.
d. Few people had any thoughts

(3) a. *Some of the boxes contain anything.
b. *Somebody talked to anybody.
c. *At least 5 students did anything.
d. *Many people had any thoughts

The notion of polarity at play here relates to a semantic notion of **monotonicity**.[1]

The notion of **monotonicity** builds on logical entailment. Monotonicity of a linguistic environment defines its entailment patterns. In (4), the domain in square brackets is **upward-entailing** (UE), or upward-monotone, – as evidenced by the valid inference from sets (*textbooks*) to supersets (*books*): sentence (4-b) entails sentence (4-a).

(4) a. Some boxes [contain books]$_{\uparrow}$
b. Some boxes [contain textbooks]$_{\uparrow}$

In contrast, (5) shows a **downward-entailing** (DE), or downward-monotone, environment, which supports inferences from sets (*books*) to subsets (*textbooks*): (5-a) entails (5-b).

(5) a. No boxes [contain books]$_{\downarrow}$
b. No boxes [contain textbooks]$_{\downarrow}$

Not all environments are either UE or DE – some are **non-monotone**, that is, supporting neither of the inferences:

(6) a. Exactly 5 boxes [contain books]$_{-}$
b. Exactly 5 boxes [contain textbooks]$_{-}$

Expressions responsible for monotonicity of a linguistic context are a heterogeneous class that includes sentential operators such as negation and conditional *if*; quantifiers (*some*, *no*, *few*, *at most five* etc.); quantificational adverbs (*rarely*, *always* etc.) and more.

Monotonicity is a highly abstract logical property interfacing with general reasoning. At the same time, it is deeply embedded into natural language grammar and it is relevant for understanding of inner workings of different linguistic expressions, such as NPIs.

As shown by examples (1)-(3), DE contexts give rise to negative polarity, as seen from NPI acceptability; UE contexts are positive. There is conflicting evidence concerning non-monotone contexts (Crnič, 2014; Alexandropoulou et al., 2020).

The connection between monotonicity and NPI licensing is undeniable also beyond examples (1)-(3) (see Fauconnier 1975; Ladusaw 1979 and much

[1]This is a simplification. This is true of so-called 'weak NPIs' – a subclass of NPIs to which *any* belongs. We will keep referring to them simply as NPIs since we are only discussing weak ones. There are also other factors in weak NPI distribution apart from monotonicity (see Giannakidou 1998; Barker 2018). Still, we focus on monotonicity as a crucial factor in NPI acceptability, following evidence discussed in the rest of the section.

Logical monotonicity		Subjective monotonicity	
NEG >> AFF;	AT MOST > AT LEAST	NEG > AT MOST;	NO > FEW
NO >> SOME;	AT MOST > BETWEEN / EXACTLY	NEG > FEW;	NO > FEWER
FEW > MANY;	FEW > BETWEEN / EXACTLY	NEG > FEWER;	FEWER > AT MOST
FEWER > MORE;	FEWER > BETWEEN / EXACTLY	NO > AT MOST;	EXACTLY > BETWEEN

Table 1: Graded monotonicity: summary of psycholinguistic experimental results (Geurts, 2003; Sanford et al., 2007; Chemla et al., 2011; McNabb et al., 2016; Denić et al., 2020). The order in pairs represents that the first element is judged as a better NPI licenser than the second one or that it better supports DE inferences (or both). That is, 'NEG >> AFF' reads as 'Sentences with sentential negation show much higher level of NPI acceptability or support DE inferences more than simple affirmative sentences.'. The 'Logical monotonicity' side of the table groups together all relations expected under the logical view of monotonicity; 'Subjective monotonicity' contains additional asymmetries found experimentally that do not follow from the simple logical view.

subsequent literature). Experimental evidence shows a bi-directional connection between inference judgments in a context and NPI acceptability in that context. Chemla et al. (2011) found that the inferences a person considers valid in a given linguistic context predict how acceptable they would find an NPI in that same context. Conversely, Denić et al. (2020) show that inferential judgments are modified by the presence of an NPI. So, the two phenomena show clear mutual influence.

Importantly, both monotonicity and NPI acceptability in humans is not an all-or-nothing matter. Acceptance of logically valid inferences and rejection of invalid ones varies to some extent from person to person – and from context to context (Geurts, 2003; Sanford et al., 2007; Chemla et al., 2011; McNabb et al., 2016; Denić et al., 2020).

Chemla et al. (2011) report that logically DE sentences with *no* are perceived as DE by human participants only 72% of the time. *At most* – also logically a DE environment – is only recognized as such 56% of the time. Moreover, *less than* and *at most* – truth-conditionally equivalent environments – differ in DE inference endorsement by 11%. The best predictor of NPI acceptability by humans was found to be not the logical entailment pattern but the subjective, or perceived, one (Chemla et al., 2011; Denić et al., 2020).

There is no single overarching psycholinguistic study testing the whole landscape of contexts. Combined knowledge from an array of studies (Geurts, 2003; Sanford et al., 2007; Chemla et al., 2011; McNabb et al., 2016; Denić et al., 2020) produces the picture summarized in Table 1.

3 Previous work

NPIs have been a topic of an investigation in the context of LRMs, both as a subset of a more general test dataset (Marvin and Linzen, 2018; Hu et al., 2020), and as the main object of study (Jumelet and Hupkes, 2018; Warstadt et al., 2019; Jumelet et al., 2021; Weber et al., 2021). Here we focus on (Warstadt et al., 2019) as a representative case, as it shares with other previous studies its general set-up: assessment of LRMs against predictions of linguistic theory.

Warstadt et al. (2019) focus on NPIs in BERT. Using a variety of testing techniques, both zero-shot and with fine-tuning, they conclude that BERT's ability to recognize NPI licensing environments and, therefore, to tell licit uses of NPIs from illicit ones varies a lot depending on the type of context, scope configuration and the type of experimental setting.

This might lead one to conclude that BERT's ability to recognize polarity of a sentence is not so great across the board. Indeed, reports from other tasks that involve polarity and/or monotonicity seem to support this. In particular, natural language inference has been reported to be hard for LRMs (Yanaka et al., 2019a,b; Talmor et al., 2020; Geiger et al., 2020). Remarkably, Geiger et al. (2020) report that fine-tuning BERT on the SNLI dataset and then evaluating it on DE sentences (their NMoNLI dataset) results in 2.2% accuracy – that is, the model practically ignores the monotonicity profile of the sentence. But is alleged poor polarity detection to blame here?

Importantly for our study, Warstadt et al. (2019) judge BERT's recognition of NPI acceptability against logical monotonicity rather than subjective monotonicity as uncovered by psycholinguistic experiments. So, we believe that these results deserve a second look.

One of the measuring techniques in Warstadt et al. (2019) is very close to one of the two tech-

niques we will adopt in this paper. It is a version of Cloze Test adapted for MLM, where probabilities of candidates for the masked position are compared. We discuss the set-up in section 4.

Finally, the idea of targeted LRM evaluations modeled after **psycholinguistic experiments** is being used in an increasing number of recent studies, albeit mainly in the domains of syntax and lexical semantics (Gulordava et al., 2018; Linzen et al., 2016; Marvin and Linzen, 2018; Wilcox et al., 2018; Chowdhury and Zamparelli, 2018; Futrell et al., 2019; Nair et al., 2020; Abdou et al., 2020; Ettinger, 2020).

We move on to describing our dataset, procedure and results.

4 Method

We perform two types of tests using the dataset that we produce for this purpose. One experiment is done with BERT, the other one with GPT-2. Both experiments are performed in a zero-shot setting – using the pre-trained models without fine-tuning. The goal of these experiments is to test the contrasts between types of sentences described in Table 1. We will do this by comparing the relevant pairs of contexts along LRM-derived metrics that are meant to capture grammaticality / acceptability.

First, we describe the dataset; then we explain the experiment procedure for BERT and GPT-2; finally, we report and discuss the results.

4.1 The ANY dataset[2]

Our dataset consists of two parts: one with natural and one with synthetic data.

4.1.1 Natural data

We scraped the Gutenberg Project and a subset of English Wikipedia to obtain the list of sentences that contain *any*. Next, using a combination of heuristics[3], we filtered the result with regular expressions to produce two sets of sentences (the second set underwent additional manual filtering):

- 3844 sentences with sentential negation and a plural object with *any* to the right to the verb;
- 330 sentences with *nobody / no one* as subject and a plural object with *any* to the right.

The first set was modified to substitute the negated verb by its non-negated version, so we contrast 3844 sentences with negation and 3844 affirmative ones (NEG vs. AFF). In the second dataset, we substituted *nobody* for *somebody* and *no one* for *someone*, to check the SOME vs. NO contrast.

4.1.2 Synthetic data

We used the following procedure. First, we automatically identified the set of verbs and nouns to build our items from. To do so, we started with `bert-base-uncased`[4] vocabulary. Taking its non-subword lexical tokens is an easy way to get a list of simple and common words. We ran this list through a SpaCy POS tagger[5]. Further, we lemmatized the result using `pattern`[6] and dropped duplicates. Then, we filtered out modal verbs, singularia tantum nouns and some visible lemmatization mistakes. Finally, we filtered out non-transitive verbs to give the dataset a bit of a higher baseline of grammaticality.[7]

We kept top 100 nouns and top 100 verbs from the resulting lists – these are the lexical entries we will deal with. Then, we generated sentences with these words, using the following pattern:

A(n) $\texttt{noun}_x$ `verb.PST.SG` a(n) $\texttt{noun}_y$[8]

For this, we iterate over the 100 nouns in the subject and the object positions (excluding cases where the same noun appears in both positions) and over the 100 verbs. The procedure gave us 990k sentences like these:

(7) a. A girl crossed a road.
b. A community hosted a game.
c. A record put an air.

Some are more natural, make more sense and adhere to the verb's selectional restrictions better than the others. To control for this, we ran the sentences through `GPT-2`[9] and assigned perplexity to all candidates. Then we took the bottom 20k of the sentences (≈ the most 'natural' ones) as the core of our synthetic dataset.

[2] The data are available at https://github.com/altsoph/Transformers-in-the-loop

[3] The script that can be used to reproduce the filtering procedure is available in the project repository, see fn. 2.

[4] https://huggingface.co/bert-base-uncased

[5] https://github.com/explosion/spacy-models

[6] https://pypi.org/project/Pattern/

[7] Our procedure was equivalent to that in github.com/Mirith/Verb-categorizer

[8] We use the singular indefinite object for this part of the procedure to avoid idiomatic verb phrases (*change hands*, *join forces*) at the top of the list.

[9] https://huggingface.co/gpt2

We tried to approximate the 'naturalness' of examples by a combination of measures. We rely on insights from different models (GPT-2, BERT, corpus-based statistical insights into verb transitivity) on different stages of the dataset creation. Still, some sentences sound intuitively 'weird'. We do not see this as a problem though – we will not rely directly on the naturalness of individual examples, rather we will measure the effect of the NPI across the dataset (as is common practice when working with synthetic data – see, for example, Geiger et al. 2020, 2021). The amount of the examples will allow us to generalize across varying parts of the sentences to make sure that the results can be attributed to the parts we are interested in: items responsible for the polarity of the sentence. The quantity of test items is crucial for reproducing psycholinguistic experiments on LRMs – while in the former one sentence gives rise to a number of observations when different human participants make a judgment, in the latter one test sentence gives one observation only.

With this in mind, we use the 20k sentences produced by the previous steps to build the parts of our synthetic dataset. Each of the sentences has a pluralized (not singular anymore!) object in combination with *any*: `any roads`. The subject type varies in different datasets comprising our synthetic data. Here is what we end up with:

- 12 datasets 20k sentences each:
  AFF (8-a); NEG (8-b); SOME (8-c); NO; MANY; FEW; MORE THAN 5; FEWER THAN 5; AT LEAST 5; AT MOST 5; EXACTLY 5; BETWEEN 5 AND 10;

- 2 datasets 8230 sentences each:
  SOMEBODY / SOMEONE / SOMETHING (8-d); NOBODY / NO ONE / NOTHING (replacing the whole subject, duplicates deleted)

(8) a. A girl crossed any roads.
b. A girl didn't cross any roads.
c. Some girls crossed any roads.
d. Somebody crossed any roads.

Overall, sentences in all parts of our dataset vary in the type of context it instantiates (simple affirmative, negation, different quantifiers) – but all sentences contain *any* in the object position in combination with a plural noun.

The next two subsections explain the metrics derived from the two model we study, stemming from the differences in their architecture and training objectives.

4.2 BERT: Cloze Test

The Cloze Test on BERT is very similar to that described in (Warstadt et al., 2019). In each of the sentences in the dataset, we mask *any* and ask BERT for predictions for the masked position:

```
[CLS] Few girls crossed [MASK] roads . [SEP]
```

We extract the probability that BERT assigns to *any* in the masked position, as well as the rank of *any* in BERT vocabulary sorted by the probability in the masked position.

Further, we compare these values between conditions (= different types of contexts). The comparison between a pair of conditions will be expressed as the percentage of sentences in our dataset where *any* got a higher probability in the first condition compared to the probability of *any* in the corresponding sentence in the second condition. The same for the rank of *any* instead of probability. For example, $\langle$AFF: NEG$\rangle$: 0.12% reads as: in 0.12% of the dataset, *any* got a higher probability (or a higher rank) in an affirmative sentence compared to the corresponding sentence with negation. Intuitively: that most of the time, a sentence with negation makes a better environment for *any* than the minimally different affirmative sentence.

4.3 GPT-2: Perplexity difference

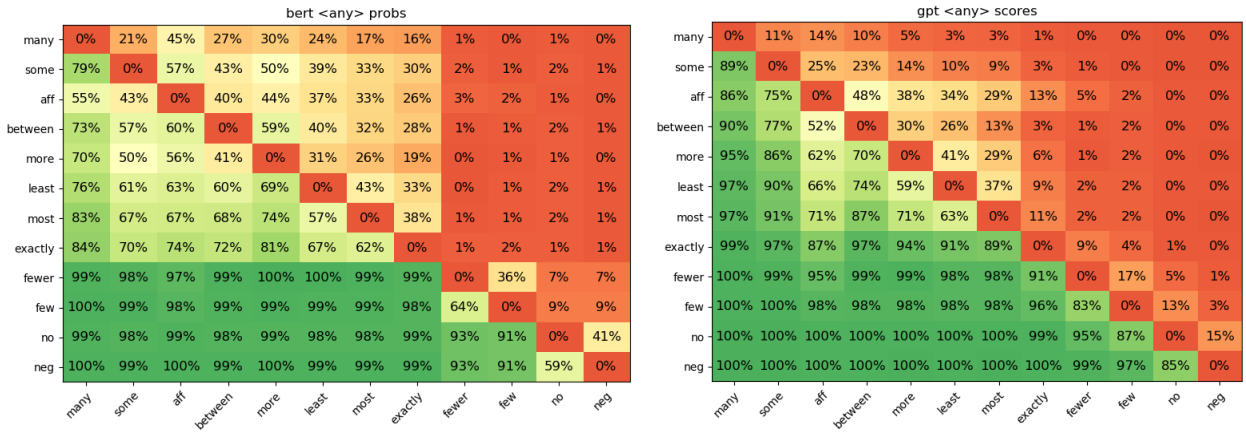
In this test, for each sentence in the dataset, we calculate perplexity of this sentence (9-a) according to the GPT-2 model – and perplexity of that same sentence with *any* deleted (9-b):

(9) a. Few girls crossed any roads.
b. Few girls crossed roads.

We take the difference between these perplexity values normalized by the number of tokens as our measure of how much the presence of *any* affects the 'naturalness' of each particular sentence.

As before, we compare these values for different conditions. For example, $\langle$AFF: NEG$\rangle$: 0.25% reads as: in 0.25% of sentences, the presence of *any* leads to a smaller increase in perplexity for the affirmative sentence, compared to the analogous negative sentence. That is, most of the time the presence of *any* worsens affirmative sentences a lot, while the corresponding negative one – less so.

This is the closest possible LM analogue of the acceptability judgment experiments like (Alexan-



(a) BERT-prob comparison across conditions

(b) GPT-PPL-diff comparison across conditions

Figure 1: LRM experiment results

dropoulou et al., 2020), which measure the differences between acceptability scores with and without *any* for different types of contexts.

5 Results of model evaluation

We will discuss results from BERT and GPT-2 together, because they mostly agree.

One general result that allows us to limit our attention to one of the two BERT metrics is that BERT rank and BERT probability produce the same order on all condition pairs of interest except for one (⟨AT MOST, AT LEAST⟩) and we will only discuss BERT probabilities in this section.

The 20k synthetic data results are summarized in Fig. 1. The conditions in the 20k results are sorted for readability. 8k synthetic data results: ⟨NO-, SOME-⟩: 99.76% (BERT-prob); 99.56% (GPT-PPL-diff).

In short, **all predictions based on psycholinguistic evidence discussed in section 2 (Table 1) are confirmed by our LRM data.**

As a sanity check, we compare these results with the results of the same procedure on our natural dataset, and they are very similar: ⟨NEG, AFF⟩: 97.21% (BERT-prob), 97.17% (GPT-PPL-diff); ⟨NO-, SOME-⟩: 98.29% (BERT-prob), 96.98% (GPT-PPL-diff).

The take home message from these results is that **LRMs can tell between negative and positive polarity, as well as between different types of contexts by their monotonicity, as measured by NPI acceptability**. Moreover, what is encoded is a subjective version of the relevant property, similar to what is reflected in graded non-categorical judgments seen in psycholinguistic experiments.

Establishing this, first of all, helps us make more sense of the metrics derived from such models and helps draw a more accurate line between noise and meaningful output. Second, it encourages a closer tie between experiments with humans and with LRMs: LRMs encode a snapshot of numerous subjective linguistic intuitions, and maybe we can use LRMs to get indirect access to speakers' shared intuitions as a source of new theoretically relevant linguistic generalisations. The next section is a pilot attempt in this direction. We establish a new generalization looking at LRM data – and then confirm it in a psycholinguistic experiment.

6 Next step: Cardinality dependency

For the conditions which involve numerals we left one parameter unexplored so far, namely, the numeral itself. In this section, we look at the dependency between NPI acceptability and the numeral.

There is no experimental data on this. Theoretical literature tentatively suggests that the higher the numeral, the less acceptable an NPI in its scope (Crnič, 2014):

(10) Exactly two of the boxes contain anything
(11) ??Exactly 98 of the boxes contain anything

However, the judgments are subtle and theoretical discussion still waits for an empirical basis. Let us look at our conditions with numerals (apart from BETWEEN – we set it aside as too complicated). For each of the conditions, we keep everything constant apart from the numeral and check the effect the numeral has on NPI acceptability.

6.1 As seen in LRMs

We looked at numerals with these numeric values: $[2-20, 30, 40, 50, 60, 70, 80, 90]$. As before, we made pair-wise comparisons between sentences in our synthetic dataset that differ only in the numeral it contains. The measures are the same as before.

Both models show an upward trend: the higher the numeral, the worse the context becomes for *any*. This tendency is shown on Fig. 2.

The lines show comparison between sentence pairs in which the second one has a numeral higher than the one in the first sentence by n, where n is plotted on the x axis (so, 10 on the x axis comprises all pairs that differ by 10 – $\langle 2, 12\rangle$,$\langle 3, 13\rangle$...). On y, we show the percentage of pairs in which the first sentence showed higher probability of *any* than the second one.

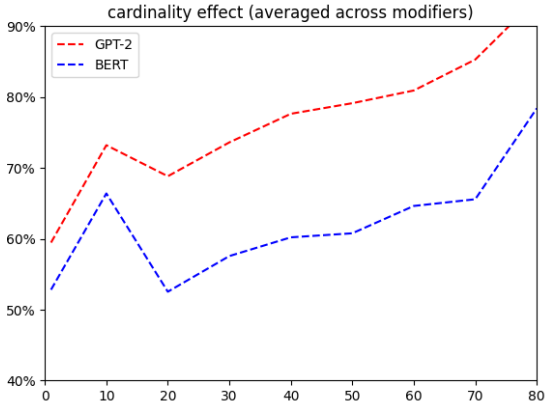


Figure 2: The effect of numeral on `any`.

The effect of the numeral on the NPI acceptability can be sometimes quite strong: to the point of flipping the 'better NPI licenser' relation in a pair of contexts. For example, this is the case for AT LEAST and MORE THAN in BERT. They have the same logical monotonicity profile (both UE). However, we can find a pair of numerals such that flipping them orders the resulting contexts differently:

AT LEAST 2 > MORE THAN 70: 94%
MORE THAN 2 > AT LEAST 70: 68%

Let us check the effect of numeral on humans, as well as a licensing flip due to the numeral.

6.2 In humans

For the ease of comparison between our LRM experiment data in the previous section and the experiment on human participants, we formulate the latter as a **forced-choice task**.

The participants saw pairs of sentences and were instructed to pick the one that is more grammatical. The study has a **2x2 design** with these factors:

- NUMERAL: *five* vs. *seventy*
- QUANTIFIER: *at least* vs. *more than*

This gives six forced-choice test conditions:

at least five vs. *at least seventy*
at least five vs. *more than five*
at least five vs. *more than seventy*
at least seventy vs. *more than five*
at least seventy vs. *more than seventy*
more than five vs. *more than seventy*

These prefixes were used to generate pairs of sentences using patterns from the 20k synthetic dataset. We randomly selected 50 out of the 20k patterns, which results in 2500 pattern pairs. With 6 test conditions, this amounts to **15k unique test items**.

We used Yandex.Toloka to recruit self-reported native speakers of English for this experiment.[10] They were allowed to complete the full task after they passed a test with 10 control items with 7 or more correctly identified grammatical sentences.

In the main part of the task, each participant saw **38 pairs of sentences**: 22 were filler/control items and 16 test items. All participants saw the same filler/control items (random order), test items were taken from the pool of 15k test items in random order and evaluated with no overlap.

In total, 968 participants were recruited. We filtered out the data from those who gave wrong answers to more than 30% of the filter/control items in the main part of the task. We were left with 656 participants (= 10496 test items; more than a 2/3 of our pool of test items). Fig. 3 shows the **results** of the experiment. We used the binomial test to analyze the data. The boxes in the plot show the 95% confidence interval.

Result #1: The effect of the numeral is confirmed both within and across the two types of contexts (lines 1, 6, 9 and 10 in Fig. 3). **Result #2:** AT LEAST and MORE THAN are not ordered with respect to each other (lines 7 and 8). It is possible to find a particular numeral where the difference reaches significance (line 2), but overall there is no clear order. **Result #3:** Our data do not show a statistically significant flip between contexts with different numeral values. Even though one side of the flip is there (line 3), the flip of this pair did not reach significance (line 5).

[10] https://toloka.yandex.com/for-requesters

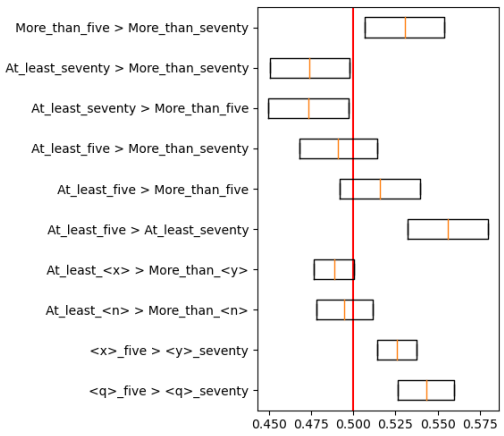


Figure 3: Human judgments of *any*-acceptability

Conclusion: The results are generally in line with the trend observed in section 6: the higher the numeral, the worse the context gets for an NPI. This is the first experimental confirmation of this effect, to the best of our knowledge. It is noteworthy that we first found it via LRM – and then confirmed it with human participants.

A more specific result of this effect – what we call a 'flip' – is seen in our data as a tendency, but the effect did not reach significance. It could be an LRM artifact – or the lack of it could be an artifact of our experiment. A different choice of numerals or a higher number of participants could sharpen these results. We leave this for future work.

7 Discussion and outlook

Our experiments provide solid support for an approach under which LRM performance is compared directly to psycholinguistic data rather than to predictions of a linguistic theory. This opens up prospects for research that will result in a more empirically grounded picture of where the limits of LRM abilities lie.

Our results tell us something new about LRMs but also suggest that LRMs can be included in the experimental loop of theoretical semantics alongside with traditional experiments. To pilot this idea, we conducted an experiment on the effect of the numeral on NPI acceptability. We confirmed our LRM findings in a parallel psycholinguistic study.

In this paper, we only explore the connection between behavioral experiments and LRM-derived metrics. What about online measures in psycholinguistic studies? Can we find a usable analogue to, for example, eye-tracking or reaction times in self-paced reading studies – that is, studies that tell us which parts of input are important in processing? One obvious LRM-based candidate is attention.

We took a preliminary look at BERT attention distribution in sentences with *any* in an attempt to identify the attention head that contributes most to monotonicity-via-NPIs (see Voita et al. 2019 for a discussion of attention head specialization). To factor out linear position, we focused on the natural part of our dataset. We took the sentences that contained both a quantifier with a clear monotonicity profile (*somebody*, *nobody*, *someone* etc.) and *any*; calculated attention from *any* to the quantifier for every layer and every attention head and averaged it across sentences. Then we sorted the results and went through the top of the resulting list.

We found that the attention head (6,2) of `bert-base-uncased` model – 6th layer, attention head 2 – seems to specialize in precisely what we are looking for. Saliency maps below show that in a variety of contexts beyond the ones we checked for the purposes of this paper, monotonicity-affecting items are highlighted – buttressing the hypothesis that monotonicity is important for NPI licensing (*without*, *do*-support in a question, *if*, lexical negation):

[CLS] it felt odd without any wards on it . [SEP]

[CLS] do you have any brothers or sisters ? [SEP]

[CLS] if there ' d been any babies present , he ' d have been un ##sto ##ppa ##ble . [SEP]

[CLS] we are unable to identify any others who knew of the scheme at the time it was being considered . [SEP]

Additionally, this attention head reflects the role of the numeral in NPI licensing that we established in section 6: in all contexts with numerals that we looked at, a lot of attention goes from *any* to both the quantifier (say, *exactly*) and the numeral that comes with it. Moreover, the higher the numeral, the more attention goes to it, compared to the amount of attention that goes to the quantifier:

[CLS] exactly two games told any stories . [SEP]

[CLS] exactly ninety games told any stories . [SEP]

More work is needed to verify and interpret these patterns systematically and compare them to other attribution measures and to online metrics in psycholinguistic studies.

Acknowledgements

We thank the anonymous ARR reviewers; the audience and organizers of the CNRS Seminar on the Interactions between Formal and Computational Linguistics; Yandex.Toloka for the help with the human assessment study. We also thank Alexandre Cremers, Ekaterina Garmash, Borislav Kozlovskii, Rick Nouwen, and Denis Paperno for the discussions of our ideas and earlier versions of the paper.

References

Mostafa Abdou, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott, and Anders Søgaard. 2020. The sensitivity of language models and humans to Winograd schema perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7590–7604.

Stavroula Alexandropoulou, Lisa Bylinina, and Rick Nouwen. 2020. Is there 'any' licensing in non-DE contexts? An experimental study. In *Proceedings of Sinn und Bedeutung*, volume 24, pages 35–47.

Chris Barker. 2018. Negative polarity as scope marking. *Linguistics and philosophy*, 41(5):483–510.

Marco Baroni. 2021. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *arXiv preprint arXiv:2106.08694*.

Emmanuel Chemla, Vincent Homer, and Daniel Rothschild. 2011. Modularity and intuitions in formal semantics: The case of polarity items. *Linguistics and Philosophy*, 34(6):537–570.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th international conference on computational linguistics*, pages 133–144.

Luka Crnič. 2014. Non-monotonicity in NPI licensing. *Natural Language Semantics*, 22(2):169–217.

Milica Denić, Vincent Homer, Daniel Rothschild, and Emmanuel Chemla. 2020. The influence of polarity items on inferential judgments. Submitted.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. volume 8, pages 34–48. MIT Press.

Gilles Fauconnier. 1975. Polarity and the scale principle. In *Proceedings of Chicago Linguistc Society 11*, pages 188–99.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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 32–42.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *arXiv preprint arXiv:2106.02997*.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173.

Bart Geurts. 2003. Reasoning with quantifiers. *Cognition*, 86(3):223–251.

Anastasia Giannakidou. 1998. *Polarity sensitivity as (non) veridical dependency*, volume 23. John Benjamins Publishing.

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

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.

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.

Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess NPI licensing. *CoRR*, abs/2105.13818.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? on the ability of lstms to understand negative polarity items. In *BlackboxNLP@ EMNLP*.

William A Ladusaw. 1979. *Polarity sensitivity as inherent scope relations*. Ph.D. thesis, Austin, TX: University of Texas at Austin.

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

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 521–535. MIT Press.

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.

Yaron McNabb, Stavroula Alexandropoulou, Dominique Blok, Sofia Bimpikou, and Rick Nouwen. 2016. The likelihood of upper-bound construals among numeral modifiers. In *Proceedings of Sinn und Bedeutung*, volume 20, pages 497–514.

Sathvik Nair, Mahesh Srinivasan, and Stephan Meylan. 2020. Contextualized word embeddings encode aspects of human-like word sense knowledge. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 129–141.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Anthony J Sanford, Eugene J Dawydiak, and Linda M Moxey. 2007. A unified account of quantifer perspective effects in discourse. *Discourse Processes*, 44(1):1–32.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. In *Transactions of the Association for Computational Linguistics*, volume 8, pages 743–758. MIT Press.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, et al. 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 (EMNLP-IJCNLP)*, pages 2877–2887.

Lucas Weber, Jaap Jumelet, Elia Bruni, and Dieuwke Hupkes. 2021. Language modelling as a multi-task problem. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2049–2060, Online. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler-gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221.

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

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019b. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 250–255.