The driving forces of polarity-sensitivity: Experiments with multilingual pre-trained neural language models

Lisa Bylinina (bylinina@gmail.com)[1]
Bookarang, Amsterdam

Alexey Tikhonov (altsoph@gmail.com)[1]
Independent researcher, Berlin

Abstract

Polarity-sensitivity is a typologically general linguistic phenomenon. We focus on negative polarity items (NPIs, e.g. English *any*) – expressions that are licensed only in negative contexts. The relevant notion of 'negative context' could be defined lexically, syntactically or semantically. There is psycholinguistic evidence in favour of semantics as a driving factor for some NPIs in a couple of languages (Chemla, Homer, & Rothschild, 2011; Denić, Homer, Rothschild, & Chemla, 2021). Testing the scale of this analysis as a potential cross-linguistic universal experimentally is extremely hard. We turn to recent multilingual pre-trained language models – multilingual BERT (Devlin, Chang, Lee, & Toutanova, 2018) and XLM-RoBERTa (Conneau et al., 2019) – and evaluate the models' recognition of polarity-sensitivity and its cross-lingual generality. Further, using the artificial language learning paradigm, we look for the connection in neural language models between semantic profiles of expressions and their ability to license NPIs. We find evidence for such connection for negation but not for other items we study.

Keywords: neural language models; polarity-sensitivity; linguistic typology; artificial language learning

Introduction

Systematic interactions between linguistic properties constrain the space of natural languages in a principled way. Studying these linguistic universals is indispensable to make progress on questions like, What are the limits of cross-linguistic variation? What is the space of possible natural languages, the space of impossible natural languages, and how do they differ from each other?

One type of universals is of special interest: those where one property semantic property and the other one is distributional. Those generalizations are particularly hard to verify, since they require a semantic account of items of interest and a way to distinguish a semantic condition from a more superficial (say, lexical) one.

We focus on one interaction relating semantics and distribution: interaction between the semantic property of **monotonicity** and a distributional property of **polarity sensitivity**.

We will focus on one type of polatity-sensitive items: **negative polarity items** (NPIs). NPIs are ungrammatical in a sentence unless the sentence provides a 'negative' context. E.g., a sentence with sentential negation like (1-a) is grammatical with *any*, while its minimal affirmative pair is not:

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

What are NPIs like *any* sensitive to? Is their grammaticality conditioned by the presence of specific words in specific positions (*not* and other 'licensers' that we'll discuss in Section 2)? Or are they actually sensitive to the meaning that these words bring into the sentence?

We will study this interaction in two large pre-trained **multilingual language models** (LMs): mBERT (multilingual BERT, Devlin et al., 2018) and XLMR (XLM-RoBERTa, Conneau et al., 2019). Our conclusions can be important for two reasons: 1) if we look at LMs as 'algorithmic theories' of the data they were trained on (Baroni, 2021), it's informative for linguistic theory to know what kind of 'theories' of polarity-related phenomena arise in LMs trained on data from 100+ languages; 2) we gain new understanding of the inner workings of multilingual LMs and the inter-language generality of their polarity representations.

This paper presents results of 4 experiments[2] that involve mBERT and XLMR (see Fig. 1):

1) We evaluate within-language polarity-sensitivity for NPIs in 4 languages (English, French, Russian and Turkish).

2) Using a simple cross-lingual transfer test, we evaluate the generality of polarity representation across languages.

3) Building on the artificial language learning paradigm (Friederici, Steinhauer, & Pfeifer, 2002; Culbertson, Smolensky, & Legendre, 2012 a.o.), we introduce a system of new tokens and acquaint the models with their monotonicity profiles. We then test whether this information is enough to trigger NPI-licensing behaviour – that is, whether NPI distribution is meaning-conditioned.

4) Finally, in an experiment parallel to Exp. 2, we conduct a cross-lingual transfer test to check whether the tokens trained to develop particular meaning profiles **on English data** interact with NPIs **in other languages** – thus checking whether cross-lingual generality of polarity encoding is a meaning-related generalization.

We start with the discussion of background and related work and then describe the experiments and their results. We report meaning-conditioning for one of our items.

[1] Equal contribution.

[2] Code and data: `https://github.com/altsoph/cogsci2022`

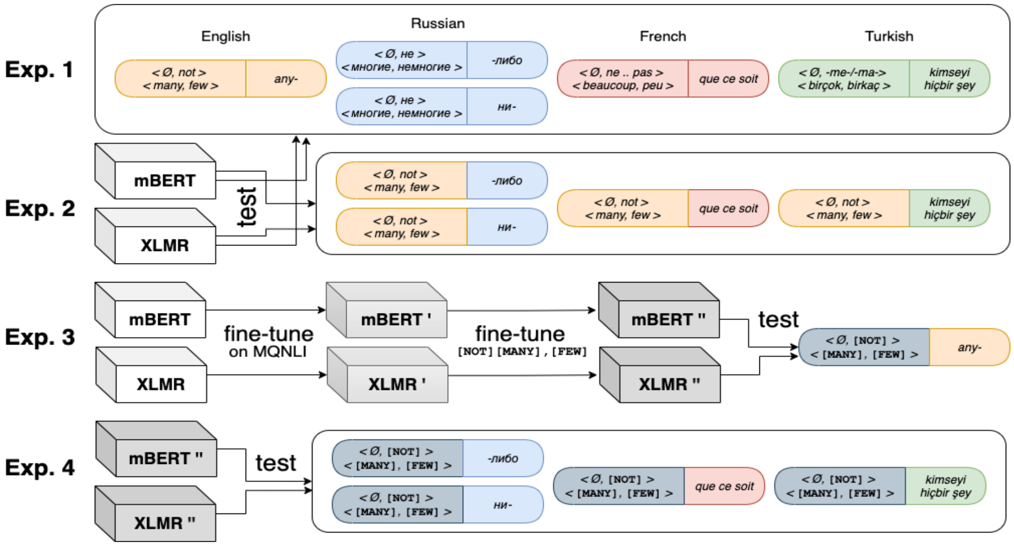

Figure 1: Experiments represented schematically. **Exp. 1** tests polarity-sensitivity in pre-trained mBERT and XLMR for 4 languages individually. **Exp. 2** tests cross-lingual transfer of polarity sensitivity with hybrid sentences where English polarity operators (*not*, *many*, *few*) are transplanted into sentences from different languages containing NPIs. **Exp. 3** exposes models to new tokens and their semantic profiles via NLI training in English and then tests their NPI-licensing behaviour. **Exp. 4** tests cross-lingual transfer of NPI licensing properties of new tokens, recycling Exp. 2 procedure.

Linguistic background

NPIs are expressions with limited linguistic distribution. The distribution of NPIs like *any* is more intricate than suggested by the simple minimal pair in (1). For instance, (2) are 'negative enough' to license *any*, while (3) are not – even though none contain sentential negation.

(2) a. No boxes contain anything.
b. Few people had any thoughts.

(3) a. *Some boxes contain anything.
b. *Many people had any thoughts.

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) – 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 $]_{\uparrow}$

(5) shows a **downward-entailing** (DE) environment: (5-a) entails (5-b) – an inference from sets to subsets.

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

The exact relation between NPIs and monotonicity of the context where they appear has been subject to a lot of research in linguistics (see Fauconnier, 1975; Ladusaw, 1979 and much subsequent literature). Importantly, both monotonicity and NPI grammaticality are not an all-or-nothing matter when it comes to human judgments about them (Geurts, 2003; Sanford, Dawydiak, & Moxey, 2007; Chemla et al., 2011; McNabb, Alexandropoulou, Blok, Bimpikou, & Nouwen, 2016; Denić et al., 2021). Chemla et al. (2011) report that sentences with *no* are perceived as DE by human subjects only 72% of the time. *At most* is only recognized as DE 56% of the time. This variation allows for a study that investigates correlation between NPI acceptability judgments and monotonicity judgments about the same context. Chemla et al. (2011) found that the inferences a person considers valid in a given context predict how acceptable they would find an NPI in it. Conversely, Denić et al. (2021) show that inferential judgments are modified by the presence of an NPI. That is, speakers have this correlation as part of their (implicit) linguistic knowledge: NPIs tend to occur in contexts that also tend to have some monotonicity profile, and information about one can change your expectation about the other.

There are open questions concerning this interaction. One, it still remains to be shown that this is a causal relation. Two: if it is, is it a fundamental property of natural language that can be found beyond a couple of languages? The questions of causality and cross-linguality are inter-related. A language typically only has few actively used NPIs and a limited set of monotonicity-altering expressions. This limited data is compatible with different causal analyses that can end up as linguistic knowledge of a speaker. But what kind of causal structure of this domain would one develop as a speaker of 100+ languages? Would it be the best generalization to condition your NPIs by monotonicity? Would it be a good move to generalize these phenomena across languages at all? Humans speaking 100+ languages natively are hard to find. But we do have multilingual language models.

Related work

Polarity and monotonicity in LMs. NPIs have been the topic of several studies in the context of LMs (Marvin & Linzen, 2018; Hu, Gauthier, Qian, Wilcox, & Levy, 2020; Jumelet & Hupkes, 2018; Warstadt et al., 2019; Jumelet, Denić, Szymanik, Hupkes, & Steinert-Threlkeld, 2021). A variety of techniques have been used to assess to what extent LMs recognize the relation between NPIs and different types of NPI licensers. A measuring technique we will use here has been used, for instance, in (Warstadt et al., 2019): a version of the cloze test adapted for masked LMs, where probabilities of tokens in the masked position are compared. The consensus in the literature – on monolingual data – is that recent LMs show decent knowledge of NPI licensing conditions, with variation depending on the type of the licenser (robust with negation, less robust with other licensers like *few* etc.).

Studies of entailment pattern recognition (casted as Natural Language Inference – NLI) in LMs is less optimistic about the models' abilities (Yanaka et al., 2019a, 2019b; Talmor, Elazar, Goldberg, & Berant, 2020; Geiger, Richardson, & Potts, 2020). However, training on more balanced datasets has been reported to help (Geiger et al., 2020; Geiger, Lu, Icard, & Potts, 2021).

Artificial language learning (ALL). The ALL framework is used in psycholinguistics and cognitive science to uncover causal relations in different domains (Friederici et al., 2002; Finley & Badecker, 2009; Culbertson et al., 2012; Ettlinger, Bradlow, & Wong, 2014; Kanwal, Smith, Culbertson, & Kirby, 2017; Motamedi, Schouwstra, Smith, Culbertson, & Kirby, 2019). It has three main ingredients: (1) a fragment of an artificial language: expressions that do not belong to the subjects' language; (2) training phase, where some information about the language fragment is given to the subjects; (3) testing phase, where it is checked what other knowledge, beside the provided, was inferred during training – revealing causal relations. ALL has only been used in the context of pre-trained LMs in (Thrush, Wilcox, & Levy, 2020), to the best of our knowledge – and never for multilingual LMs.

Cross-lingual transfer in multilingual LMs. Multilingual representations and their evaluation have been in the focus of the research community for a long time. For the history of cross-lingual representations, see Ruder, Vulić, and Søgaard (2019). We work with recent multilingual Transformer masked LMs (Devlin, Chang, Lee, & Toutanova, 2019; Conneau & Lample, 2019; Conneau et al., 2019; Siddhant et al., 2020), whose development was driven by the advances in transfer learning for NLP (Howard & Ruder, 2018; Devlin et al., 2019). Research evaluating and interpreting multilingual linguistic representations has produced several multilingual benchmarks, including XTREME (Hu, Ruder, et al., 2020) and XCOPA (Ponti et al., 2020). There are no multilingual datasets for polarity-sensitivity phenomena.

Experiment 1

Data

This experiment uses 5 language-specific datasets: one in English, one in French, two in Russian (for different NPIs) and one in Turkish. All five datasets contain synthetic dataproduced for each language following the same procedure. We describe the procedure for **English**, see the repository for complete specification.

First, we automatically identified the set of verbs and nouns to build our items from. To do so, we ran the CapitolWords corpus[3] through SpaCy POS tagger and dependency parser[4]. We sorted encountered nouns and verbs by frequency in the corpus and excluded words that appear in the MQNLI dataset (see Geiger et al., 2020 and Exp. 3 description). We filtered out verbs that had $> 50\%$ intransitive uses. We kept past tense forms of 300 top verbs meeting our criteria and plural forms of top 200 nouns. Then, we generated sentences with these words, using the following pattern:

Two `noun.PL` `verb.PRS.PL` `indef`.

The numeral is a measure to make sure the subject is plural and quantifiable; the present tense serves the same purpose. Finally, `indef` varies over $\{$*somebody*, *something*$\}$ to capture verbal selectional restrictions. The pattern gives 120k sentences like *Two activities mean something*.

We ran the sentences through GPT-2[5], sorted them by GPT-2 perplexity and kept the bottom 10k as the basis of the test dataset. Then we constructed 10k quadruples:

⟨AFF, NEG, MANY, FEW⟩

The sentences in the tuple are generated using simple morphosyntactic patterns (`any-` varies between *anything* and *anybody* depending on the indefinite in the original sentence):

(6) a. The letters meant anything.
b. The letters **did not** mean anything.
c. **Many** letters meant anything.
d. **Few** letters meant anything.

Here are French, Russian and Turkish examples, respectively:

(7) a. Peu d'amis ont vu quoi que ce soit.
few of.friends have seen anything
b. Люди ничего не потеряли.
people anything NEG lost
c. Doktorlar kimseyi aramadı.
doctors anybody called.NEG

Procedure

To measure polarity-sensitivity, we follow the procedure in (Warstadt et al., 2019 a.o.). We target two contrasts: ⟨AFF, NEG⟩ and ⟨MANY, FEW⟩. For each of the relevant pairs in each quadruple in each dataset, we mask NPIs with one or more `[MASK]` tokens, depending on the length of the tokenized NPI. Then, we measure the proportion of the dataset

[3] Accessed via `textacy.datasets`.

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

[5] `https://huggingface.co/gpt2`

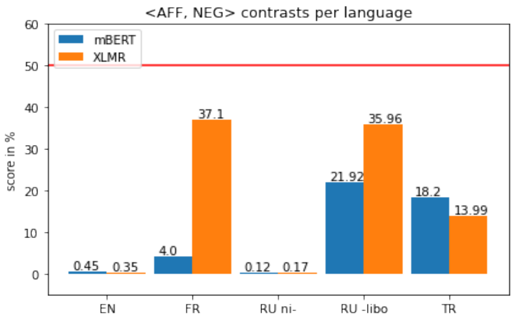

Figure 2: Results of Exp. 1 per model and language

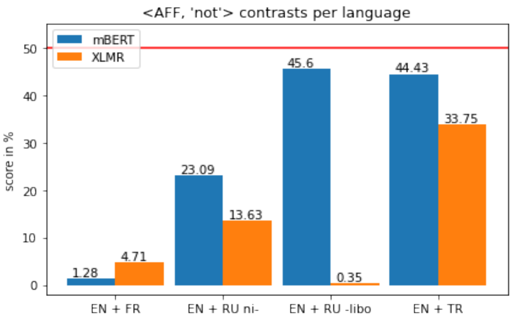

Figure 3: Results of Exp. 2 per model and language

where the probability of the NPI, according to the model, is higher in the first element of the pair. This is described in Formula (1) (where $|D|$ stands for the dataset cardinality) for AFF vs. NEG and one masked token. It is exactly the same for MANY and FEW. For multi-masked cases probabilities are averaged across masks.

$$\frac{\sum_{s\in D}[p([MASK]=m|s_{\text{AFF}}) > p([MASK]=m|s_{\text{NEG}})]}{|D|} \quad (1)$$

The lower the resulting number, the more pronounced NPI behaviour is. We perform this test on mBERT and XLMR.

Results

Results are summed up in Fig. 2, grouped by model and contrast pairs. The most robust contrasts are found between affirmative and negative sentences in English and with the Russian 'ни'-NPI in both models. In both of these cases, the NPI is a default choice with direct negation (*any* is an all-purpose NPI in English, Russian 'ни' is a 'strong NPI', – its use is mainly restricted to direct negation contexts). Russian 'либо' and French *que ce soit* are weak NPIs which have more prominent alternatives under direct negation ('ни' in Russian and *rien* / *personne* in French). mBERT and XLMR treat French *que ce soit* very differently in the ⟨AFF,NEG⟩ contrast. Speculatively, this reflects the difference in the generlization made by the models regarding the distribution of these items. Russian 'либо' does not show *any*-like generic-NPI behaviour – possibly, due to its competition with 'ни'. Results for Turkish are the least robust across context and models. The ⟨MANY,FEW⟩ contrasts are not very robust across the board. This fact is known for monolingual LMs (Jumelet & Hupkes, 2018; Warstadt et al., 2019; Hu, Gauthier, et al., 2020; Jumelet et al., 2021; Bylinina & Tikhonov, 2022).

Experiment 2

Data. We test cross-lingual transfer of polarity-related interaction. For this, we build on data from Exp. 1. We modify our 4 non-English datasets, transplanting English items (*not*, *many*, *few*) into sentences from other languages. This way we evaluate polarity-related interactions across a language boundary. Here is an example of an English-Russian hybrid:

(8) **Few** люди ничего потеряли.
few people anything lost

The **procedure** we follow for this experiment is exactly the same as in Exp. 1.

Results are summarized in Figure 3. Judging from these proportions, English negation is a good licenser for French NPIs, according to both models (in XLMR, better than the French negation itself). Less so with Russian 'ни' (note the difference between Figures 2 and 3 for 'ни'). Russian 'либо' is treated very differently by mBERT and XLMR, with negation/NPI interaction picked up across English and Russian by XLMR but not by mBERT. English-Turkish hybrid licensing didn't work well – speculatively, because English and Turkish negation are too different morphosyntactically for the models to make a generalization. Results are less pronounced for ⟨MANY, FEW⟩ than with negation across the board – which is expected given intra-language results from Exp. 1.

To sum up so far, both models show systematic reflexes of NPI licensing by negation, less so by quantifiers. There is variation between models on how weak NPIs are treated in languages with a competing strong NPI / n-word under direct negation. Cross-lingual transfer varies from language to language, tentatively depending on how structurally different the languages are.

Experiment 3

Data

We use the MQNLI[6] dataset (see Geiger et al., 2020, 2021) – an NLI dataset for English with synthetic sentences generated using a template. They vary in which quantifier combines with the subject and object (*no, some, every* or *not every*); whether an adjective modifies the subject and/or object; whether there is an adverb; presence/absence of sentential negation; lexical choice of the verbs and nouns. This gives rise to logical structures of high complexity, and learning to recognize them is a challenging task for LMs. See (a) and (b) in Table 1 for examples from MQNLI.

We divided the 500k dataset into two parts that were used in different stages of Exp. 3. Stage 1 (general NLI training)

[6] https://github.com/atticusg/MultiplyQuantifiedData

Table 1: MQNLI-based data for Experiment 3: (a) and (b) are examples used in Stage 1 (taken from MQNLI); (c) illustrates an item from Stage 2, with a new token [NOT] as negation.

	Q_s	Adj_s	N_s	Neg	Adv	V	Q_o	Adj_o	N_o	label
(a)	not every	sad	baker		fairly	admits	not every	odd	idea	entailment
	some		baker	doesn't		admit	no		idea	
(b)	every	angry	philosopher	doesn't		draw	some		doors	contradiction
	every		philosopher		honestly	draws	some	Irish	doors	
(c)	every		milkman	[NOT]	stylishly	pats	not every		helmet	entailment
	every	jealous	milkman	[NOT]		pats	some	flexible	helmet	

used unmodified MQNLI items.

For Stage 2, we focused on a subset of the full MQNLI dataset. We were interested in pairs where the quantifiers don't vary lexically pairwise across sentences: item (a) in Table 1 doesn't satisfy this constraint, item (b) does. So, sentences in these selected pairs can differ from each other only by the presence or absence of modifiers (adjectival or adverbial), the presence or absence of negation and the lexical content of non-quantificational words. These constraints help us focus on monotonicity specifically: in terms of monotonicity, *no* and *few* are the same, but a statement like `No P is Q` entails `It′s not the case that some P is Q` (thus relating `no` to `some`). `Few Ps are Q` does not entail that. Items from MQNLI where logical relation across quantifiers is introduced bring in information beyond monotonicity.

Out of this subset, we set aside 4k pairs where both sentences have *some* in the subject position; 4k pairs for subject *no*; the same for object position: 16k sentences in total. We also set aside 8k pairs where at least one of the sentences in the pair has negation. This gives 24k sentences that will be used in Step 2. In the 8k 'some'-sentences we replace *some* with `[MANY]`, which will be the new generic UE quantifier; same for 8k 'no'-sentences, where we replace *no* with `[FEW]` as a generic DE quantifier. Finally, in the 8k sentences with negation it was substituted with `[NOT]`, resulting in pairs like the (c) in Table 1.

The rest of the dataset will be used in Step 1 of Exp. 3 with further selection to avoid lexical overlap between splits, resulting in 20k training examples and 3.5k for validation+test.

After training, we will make use of the modified English dataset from Exp. 1 to evaluate the results of the training: we will replace English negation, *few* and *many* with new tokens `[NOT]`, `[FEW]` and `[MANY]`, respectively.

Procedure

First, we fine-tune mBERT and XLMR for NLI with the Stage 1 MQNLI data (MQNLI-1). We create a MQNLI-1 train/val/test split (ratio 84:8:8) with no lexical overlaps apart from quantifiers and negation.

Then we train the models further on Stage 2 MQNLI-based data (MQNLI-2). We extend the tokenizer vocabulary with 3 new special tokens: `[NOT]`, `[FEW]` and `[MANY]`, with randomly initialized embeddings. Then we fine-tune the models on MQNLI-2 with the same settings as before (except the data split ratio 80:10:10 for which we don't enforce lexical disjointness). We save the trained embeddings of the new tokens. We repeat Stage 2 40 times re-shuffling training data and starting with new random initialization, so that we end up with 40 triples of newly trained embeddings for each model. The NLI fine-tuning results in 90.07% test accuracy and 90.11% f1 score for mBERT (averaged across 40 runs). For XLMR – 88.55% and 88.01%, respectively. We conclude that the meanings were learned by the models successfully.

Finally, we add these new tokens to their respective original, non-fine-tuned models and assign them their newly trained embeddings. Then, using the pseudo-English dataset defined above, we test the 40 new sets of tokens for NPI-licensing potential they gained after training. We evaluate the distribution of their NPI-licensing scores against a random background: we replace the tokens in the licensing positions by 40 random tokens each and measure how random tokens license English NPIs in mBERT and XLMR. Then we compare the results for random tokens and for the new trained, following the same procedure as before.

Results

The resulting distributions – both trained and random – are shown in Figures 4-7. The NPI-licensing potential for new trained negations is significantly higher than that of random tokens in the negation position in both models (for significance estimation for Exp. 3 and 4 see Table 2). Trained `[FEW]` (vs. `[MANY]`) are in fact **worse** NPI licensers than randomly picked tokens in the same positions in mBERT (no difference from random in XLMR). That is, according to these data, the meaning of negation (as revealed by NLI) is a factor affecting its interaction with NPIs. Our experiment did not reveal the same for quantifier monotonicity.

Experiment 4

Data. This experiment tests cross-lingual transfer of polarity from tokens trained on English data in Exp. 3 to other languages. As in Exp. 2, we make hybrid sentences, where a licenser is transplanted into items from French, Russian, and Turkish, like in (9):

(9) `[FEW]` d'amis ont vu quoi que ce soit.
few of.friends have seen anything

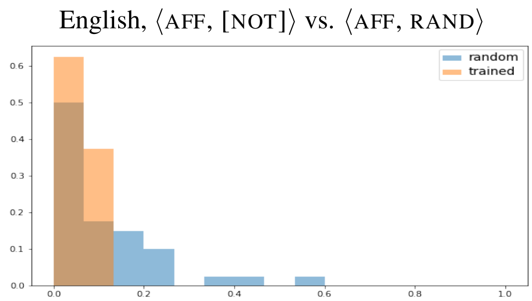
English, ⟨AFF, [NOT]⟩ vs. ⟨AFF, RAND⟩

Figure 4: mBERT (means: trained 0.056; rand 0.124)

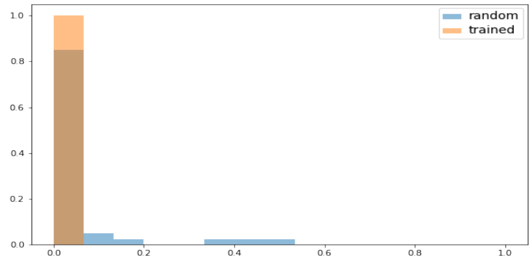

Figure 5: XLMR (means: trained 0.006; rand 0.053)

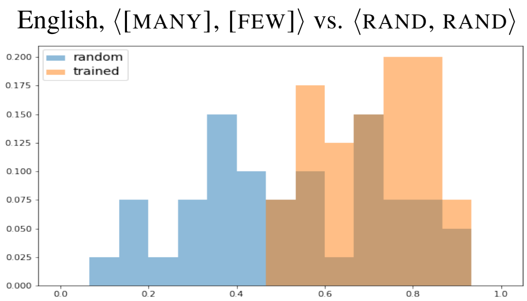
English, ⟨[MANY], [FEW]⟩ vs. ⟨RAND, RAND⟩

Figure 6: mBERT (means: trained 0.708; rand 0.518)

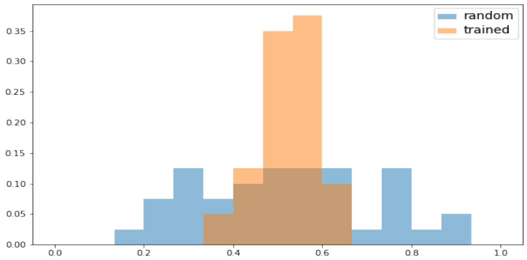

Figure 7: XLMR (means: trained 0.529; rand 0.519)

Procedure. We evaluate NPI-licensing potential of a set of new trained tokens when transplanted into non-English sentences. As a background for evaluation we use language-specific random baselines, like in Exp. 3.

Results. The only mBERT result that reached significance was the ⟨AFF, [NOT]⟩ contrast with Russian либо-NPI). XLMR showed significant levels of transfer for negation for **all languages**. The effect was not found for quantifiers (the **opposite** effect was found for Turkish in XLMR). See Table 2.

		lang	rand_mean	tr_mean	mw pval
mBERT	aff>neg	EN	0,124	0,056	0,61%
		RU ни	0,798	0,858	19,38%
		RU либо	0,86	0,823	0,073%
		FR	0,262	0,315	9,98%
		TR	0,831	0,8294	33,34%
	many>few	EN	0,518	0,708	0,01%
		RU ни	0,541	0,61	77,65%
		RU либо	0,519	0,597	21,27%
		FR	0,526	0,639	10,39%
		TR	0,552	0,488	29,19%
mBERT	aff>neg	EN	0,053	0,006	0,02%
		RU ни	0,357	0,058	0,00000008%
		RU либо	0,776	0,684	0,00000026%
		FR	0,594	0,408	0,000000001%
		TR	0,744	0,562	0,000000000%
	many>few	EN	0,519	0,529	55,07%
		RU ни	0,511	0,563	7,5%
		RU либо	0,515	0,566	13,2%
		FR	0,496	0,536	3,63%
		TR	0,48	0,563	0,02%

Table 2: Statistics and significance for Experiments 3 and 4

Discussion and conclusions

Our experiments show mixed results. First, both mBERT and XLMR systematically assign higher probability to NPIs in sentences with negation compared to affirmative sentences. However, in English, **random tokens in the position of negation also have an NPI-licensing effect**. This raises questions about the mechanisms of polarity encoding in LMs. To our knowledge, previous studies haven't used these random baselines in polarity evaluation. The contrast between UE and DE quantifiers is not very robust.

Second, cross-lingual transfer of negation from English to our target languages works some languages but not others. The reasons for the success of the transfer or lack thereof for different language pairs should be the object of a future systematic study. Notably, in XLMR, **English negation is a better licenser for French NPIs than French negation itself**; same holds for Russian 'либо'-NPIs. This means that polarity interaction in XLMR is not encoded only in the NPI, but in the negation as well. XLMR negation representations encode information on whether it's selective or not about items that can appear in its scope: English negation is omnivorous, French and Russian negations are more nuanced.

Finally, we show that negative meaning is enough for the token to develop NPI-licensing behaviour in English. Again, this behaviour of a newly introduced negative token partly transfers to other languages. We conclude that we have found some evidence in favour of cross-lingual generality in polarity-sensitivity phenomena, along with some evidence that it is related to the meaning of the NPI-licenser. We found this evidence only for negation and not for DE quantifiers: either our experiment missed it, or it is genuinely absent. We hope that our experimental set-up paves way to further studies of these phenomena in multilingual models and of other meaning-related universals.

Acknowledgements

We thank the anonymous CogSci reviewers; the audience and organizers of the CNRS Seminar on the Interactions between Formal and Computational Linguistics and of the Jabberwocky Words in Linguistics workshop. We also thank Stavroula Alexandropoulou, Jakub Dotlacil, Ekaterina Garmash, Rick Nouwen, Denis Paperno, Jan Winkowski, Yoad Winter for the discussions of our ideas and earlier versions of the paper.

References

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

Bylinina, L., & Tikhonov, A. (2022). *Transformers in the loop: Polarity in neural models of language.* arXiv. Retrieved from https://arxiv.org/abs/2109.03926 doi: 10.48550/ARXIV.2109.03926

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

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, *abs/1911.02116*. Retrieved from http://arxiv.org/abs/1911.02116

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, *32*, 7059–7069.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3), 306–329.

Denić, M., Homer, V., Rothschild, D., & Chemla, E. (2021). The influence of polarity items on inferential judgments. *Cognition*, *215*, 104791.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*. Retrieved from http://arxiv.org/abs/1810.04805

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Ettlinger, M., Bradlow, A. R., & Wong, P. C. (2014). Variability in the learning of complex morphophonology. *Applied psycholinguistics*, *35*(4), 807–831.

Fauconnier, G. (1975). Polarity and the scale principle. In *Proceedings of Chicago Linguistc Society 11* (pp. 188–99).

Finley, S., & Badecker, W. (2009). Artificial language learning and feature-based generalization. *Journal of Memory and Language*, *61*(3), 423-437. doi: https://doi.org/10.1016/j.jml.2009.05.002

Friederici, A. D., Steinhauer, K., & Pfeifer, E. (2002). Brain signatures of artificial language processing: Evidence challenging the critical period hypothesis. *Proceedings of the National Academy of Sciences*, *99*(1), 529–534.

Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. *arXiv preprint arXiv:2106.02997*.

Geiger, A., Richardson, K., & Potts, C. (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* (pp. 163–173).

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

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. (2020). A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1725–1744).

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning* (pp. 4411–4421).

Jumelet, J., Denić, M., Szymanik, J., Hupkes, D., & Steinert-Threlkeld, S. (2021). Language models use monotonicity to assess NPI licensing. *CoRR*, *abs/2105.13818*. Retrieved from https://arxiv.org/abs/2105.13818

Jumelet, J., & Hupkes, D. (2018). Do language models understand anything? on the ability of lstms to understand negative polarity items. In *BlackboxNLP @ EMNLP*.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45-52. doi: https://doi.org/10.1016/j.cognition.2017.05.001

Ladusaw, W. A. (1979). *Polarity sensitivity as inherent scope relations*. Unpublished doctoral dissertation, Austin, TX: University of Texas at Austin.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 1192–1202). Retrieved from https://www.aclweb.org/anthology/D18-1151 doi: 10.18653/v1/D18-1151

McNabb, Y., Alexandropoulou, S., Blok, D., Bimpikou, S., & Nouwen, R. (2016). The likelihood of upper-bound construals among numeral modifiers. In *Proceedings of Sinn und Bedeutung* (Vol. 20, pp. 497–514).

Motamedi, Y., Schouwstra, M., Smith, K., Culbertson, J., & Kirby, S. (2019). Evolving artificial sign languages in the lab: From improvised gesture to systematic sign. *Cognition*, *192*, 103964. doi: https://doi.org/10.1016/j.cognition.2019.05.001

Ponti, E. M., Glavaš, G., Majewska, O., Liu, Q., Vulić, I., & Korhonen, A. (2020). XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2362–2376).

Ruder, S., Vulić, I., & Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, *65*, 569–631.

Sanford, A. J., Dawydiak, E. J., & Moxey, L. M. (2007). A unified account of quantifer perspective effects in discourse. *Discourse Processes*, *44*(1), 1–32.

Siddhant, A., Bapna, A., Cao, Y., Firat, O., Chen, M., Kudugunta, S., … Wu, Y. (2020). Leveraging monolingual data with self-supervision for multilingual neural machine translation. *arXiv preprint arXiv:2005.04816*.

Talmor, A., Elazar, Y., Goldberg, Y., & Berant, J. (2020). oLMpics-on what language model pre-training captures. In *Transactions of the Association for Computational Linguistics* (Vol. 8, pp. 743–758). MIT Press.

Thrush, T., Wilcox, E., & Levy, R. (2020). Investigating novel verb learning in BERT: Selectional preference classes and alternation-based syntactic generalization. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* (pp. 265–275).

Warstadt, A., Cao, Y., Grosu, I., Peng, W., Blix, H., Nie, Y., … others (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)* (pp. 2877–2887). Retrieved from `https://www.aclweb.org/anthology/D19-1286` doi: 10.18653/v1/D19-1286

Yanaka, H., Mineshima, K., Bekki, D., Inui, K., Sekine, S., Abzianidze, L., & Bos, J. (2019a). Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 31–40). Retrieved from `https://www.aclweb.org/anthology/W19-4804` doi: 10.18653/v1/W19-4804

Yanaka, H., Mineshima, K., Bekki, D., Inui, K., Sekine, S., Abzianidze, L., & Bos, J. (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)* (pp. 250–255). Retrieved from `https://www.aclweb.org/anthology/S19-1027` doi: 10.18653/v1/S19-1027