

Responses to author

We are grateful for the reviewer's thoughtful and helpful feedback. We reprint their feedback here and intersperse our responses. For ease, our responses are highlighted.

** Starting from the observed statistical differences in valence between the different classes of triggers for NPI-anymore and NPAM, the paper reaches the conclusion that NPAM really occurs with negative affect, much more than NPI-anymore. There is, however, a logical fallacy here, because it could equally be possible that there are sentences which contain an NPI-trigger but that are actually NPAM-usages - simply that trigger does not trigger anything and the interpretation as NPAM is not dependent on the trigger. The point here is that the mere presence of a trigger does not necessarily/exclusively entail a NPI-reading.

In addition, some of the examples given by the authors as examples of NPI-usages, can also be interpreted as NPAM-instances. For instance let's take "(6) You're the only thing that makes sense *anymore*". Why would a reading as an NPAM not be possible? After all if you substitute *anymore* with *nowadays* one has a perfectly fine sentence. Or "(8) Do liberals even know what they stand for *anymore*?" Again substituting *anymore* with *nowadays* seems totally acceptable.

This brings me to the following recommendation: whether *anymore* is used as NPI or NPAM should be assessed independently of the classification of triggers, because there is no univocal relationship between triggers and (pragmatic) use. If you accept that NPAM develops out of NPI, this means that there must be some ambiguity (or 'bridging contexts') in the functioning of the underlying triggers. Such an independent assessment could take place by means of a lexical substitution with *nowadays* for instance, but it requires going back to the data (or a sample of the data, if annotation is too costly). Simply by saying the *often* does triggers are good at distinguishing NPI/NPAM is not sufficient.

We appreciate this point. It's a key point of discussion on p. 21.

We have added some text in Sec. 1.1 to clarify that the gloss 'nowadays' does not clearly distinguish NPI-*anymore* from NPAM. You can force an interpretation of 'nowadays' onto most any instance of adverbial *anymore*. As such, there is no way to reliably interpret whether an utterance actually was NPAM or NPI-*anymore* in a language user's mind. This is precisely the motivation for relying entirely on whether an NPI trigger is present or not.

To the positive, though, to the extent that authors used NPAM in environments that would also license NPI-*anymore*, it would mean that our results understate the link between NPAM and negative affect. Those instances of NPAM in NPI-licensing environments would decrease valence (etc) in those environments, making them more similar to the Negative Affect and Unlicensed categories than they would be if it were possible to perfectly isolate true NPIs.

Reply

Using 'nowadays' as a test was only one of probably other potential solutions. The point remains: if one is not able to distinguish an NPI-use from an NPAM-use, then it is simply inadequate to speak of the *relation between NPI/NPAM and valence*; what is actually assessed is the *relation between trigger categories and valence*, which may or may not correlate with the relation between NPI/NPAM and valence. This is an important distinction that only increases the validity of the results and the soundness of the argumentation, I think.

Perhaps I'm misunderstanding the following, but the argument that NPAM in NPI-licensing environments would decrease the link between NPAM and valence, that is, the link that the authors are studying, is a legitimate outcome of the study, not something that we want to avoid occurring in advance, right?

More in general, the standards in NLP-studies often require features and evaluation datasets to be annotated/validated by at least two annotators, whose (lack of) agreement is measured and reported in specific indices. For instance, annotating whether an instance is NPI or NPAM for a sample of each trigger category tweets would be such an annotation. As this paper commits to techniques from that discipline, a rigorous adoption would take into account also those aspects. I think that for this venue and for the audience of Languages a fully rigorous procedure might not be a must, but I think it's important to point out the methodological implications of applying NLP-techniques.

If the position is that even a linguist should not attempt to reliably annotate these tweets—on the grounds that NPAMs cannot be distinguished from NPIs—then it raises questions about the broader meaning of the research. Such a stance invites a more fundamental discussion about the phenomena under investigation. Should there be reluctance to revise certain aspects of the analysis, a possible alternative would be to focus more on textual redaction and to clearly hedge throughout the manuscript that the investigation concerns the relationship between triggers and valence, rather than the relations between NPAM/NPI and valence.

** I feel that the classes of triggers mix syntactic with semantic information, and coupled with the fact that the study is assessing pragmatic/semantic features on these classes, makes things more unclear.

There is, for instance, a problem of circularity if one is to a-priori distinguish classes of negative words (from other, neutral words) and then verify whether the valence of these classes differ. That the class of "Negative Affect" is the most negative, seems pretty obvious. What would make more sense is to 1) define the trigger classes exclusively in syntactic terms, 2) see how much each class is associated with NPI/NPAM and 3) see whether valence/dominance/arousal differ between those classes for each of NPI/NPAM. I think that separating those three dimensions in a more principled way will offer more benefits to the analysis (which arguably obfuscates syntactic and semantic differences).

Furthermore, it is not clear to me how the class of "Semantic NEG" differs from the class of "Negative Affect" and how one the valence between those classes can be so different. It might be good to give examples of items that end up in each class, to understand if such a distinction is actually relevant.

We've revised our summary introductions of Semantic NEG and Negative Affect and hope the distinction here is clearer. We also hope this resolves some of the reviewer's other concerns about coding, as these may be a result of insufficiently clear presentation of categories. If it doesn't resolve remaining issues, I regret that I don't entirely follow the recommendation. If we separate classes entirely in syntactic terms, then our current categories of Negative Affect and Unlicensed would be combined and we'd just inflate the association between NPAM and low valence. We have made our categories intentionally conservative.

Reply

The problem does not reside in the similarity between the ‘Negative Affect’ and ‘Unlicensed’ class, but between ‘Negative Affect’ and ‘Semantic NEG’. When reading the descriptions of the classes (“obviously negatively emotionally charged adjectives and verbs” and “[...] word functions as an NPI trigger by carrying an inherently negative meaning”) I’m not so convinced there is a substantial difference that warrants the separation in two distinct classes (unless one puts it in terms of part of speech, but that seems irrelevant; or between ‘obviously’ and ‘inherently’?), definitely not on the basis of the definitions that are given in the paper.

** It feels counterintuitive to have the analysis based on the Osgood-variables come *after* the analysis based on the VADER-algorithm, because the former is clearly more coarse grained than the latter one. So you could have the Osgood-analysis as a sort of prior thing to look at, while the VADER-analysis tackles the analysis at the right level of granularity. If I can be even stronger, the Osgood-analysis doesn’t offer any benefit on top of the VADER-analysis.

Earlier on in the paper the authors argues, with reason, that individual social media posts lack context. I find that a strange rhetorical stance, given that what you do with sentiment analysis definitely does not take broader context into account (it is a typical 'reductionist' NLP-technique). The Osgood-way is even more acontextual and coarse-grained.

We haven’t taken this feedback on board. We hope we have made a case that VADER and Osgood analysis offer complementary approaches which lead us to roughly the same conclusion that sentiment analytic methods increase confidence in the link between NPAM and negative affect.

Reply

I can understand that there might be reluctance in removing the coarse-grained analysis. In that case you could simply swap the order of presentation and first discuss the Osgood-analysis and then the VADER-analysis. This seems a small and feasible change that would much improve the progression and coherence of the text.

** How come there is no multiple regression analysis on those indices that includes the interaction of dialect region and the trigger classes? It would make a lot of sense to check the impact of those two dependent variables simultaneously in one model (per index), instead of separate analyses (by means of KW-tests).

The interaction between dialect regions and trigger classes is the focus of the previous use of this dataset, which we reference frequently in this paper. We hope we have summarised these findings sufficiently to allow the current paper to stand alone without readers looking at the earlier project, but in a way that generally avoids repeating results from previous work.

Reply

Okay.