Probing word and sentence embeddings for long-distance dependencies effects in French and English

Paola Merlo
University of Geneva
Paola.Merlo@unige.ch

Abstract

The recent wide-spread and strong interest in RNNs has spurred detailed investigations of the distributed representations they generate and specifically if they exhibit properties similar to those characterising human languages. Results are at present inconclusive. In this paper, we extend previous work on long-distance dependencies in three ways. We manipulate word embeddings to translate them in a space that is attuned to the linguistic properties under study. We extend the work to sentence embeddings and to new languages. We confirm previous negative results: word embeddings and sentence embeddings do not unequivocally encode fine-grained linguistic properties of long-distance dependencies.

1 Introduction

The recent wide-spread and strong interest in RNNs has spurred detailed investigations of the distributed representations they use, learn and generate and specifically if they exhibit properties similar to those characterising human languages. For a survey see Belinkov and Glass (2019).

Results are at present rather inconclusive on whether RNNs and the representations they learn have human-like properties. While many pieces of work seems to indicate that they do, some other pieces of work have mixed results, and a few appear to show that the representations of RNNs do not match those predicted by linguistic theory or human experiments. For example, one line of work aims to correlate RNN-induced representations to linguistic properties, namely the fact that subject-verb number agreement is structure-dependent. Initial work had shown RNNs do not really learn the structure-dependency of this construction (Linzen et al., 2016), but follow up work has shown that stronger techniques can yield more positive results (Gulordava et al., 2018), only to be very promptly rebutted by work suggesting that the apparently positive results could be the artifact of a much simpler strategy, which takes advantage of the unnaturally simple structure of the examples and simply learns properties of the first word in the sentence (Kuncoro et al., 2018). Recent work by Lakretz et al. (2019), however, studies RNNs in more detail, looking at single neurons, and finds that individual neurons encode linguistically meaningful features very saliently and with behaviour over time that corresponds to the expected propagation of subject-verb number agreement information.

Similarly, probing different aspects of long-distance dependencies, so far divergent results have been reported on these constructions. While some experiments have shown that RNNs can learn the main descriptive properties of long-distance dependencies in English, for example the fact that they obey a uniqueness constraint (only one gap per filler) and also that they obey island constraints (Wilcox et al., 2018), work attempting to replicate finer-grained human judgments for French have failed to show a correlation with human behaviour (Merlo and Ackermann, 2018), while other work on English has found mixed results (Chowdhury and Zamparelli, 2018).

In this paper, we extend previous work on long-distance dependencies to tease apart the potential grounds for the different outcomes by making previous work more comparable. There are several differences between the pieces of work on long-distance dependencies mentioned above. First, the work that does not find a correspondence between the two sources of information being compared (Merlo and Ackermann, 2018) imposes a much stricter test of correspondence — total correlation— than the general effect reported in Wilcox et al. (2018). Secondly, the pieces of work vary in task: it is possible that word embed-

Proceedings of the Second BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 158–172
Florence, Italy, August 1, 2019.

dings need to be used holistically in a prediction task similar to what humans solve to show a positive correlation. Finally, these pieces of work are on different languages.

Beside these differences in experimental setup, it is also possible that holistic representations such as word embeddings need to be transformed and translated into the right space to show correlations with human judgments. Specifically, word embeddings are a merger of the many levels of representation that we find in human languages: lexical, morphological, syntactic, semantic. It has been argued that post-processing transformations can tease apart syntactic aspects of distributed representations from semantic aspects (Artetxe et al., 2018).

Based on all these observation, we extend the work of Merlo and Ackermann (2018), which had found no correlation, along these lines. To preview, while we are able to get slightly better correlations to human judgments than those reported by Merlo and Ackermann (2018), the mixed results are confirmed: word embeddings and sentence embeddings, the representations produced by RNNs, do not unequivocally encode fine-grained linguistic properties of long-distance dependencies.

2 Intervention effects in human sentence processing

A core distinguishing property of human languages is the ability to interpret discontinuous elements as if they were a single element. These are called long-distance dependencies.

For example, sentence (1a) is an object-oriented restrictive relative clause, where the object of the verb phrase *annoying* is also the semantic recipient of the verb *smile*, connecting two distant elements. Long-distance dependencies are not all equally acceptable or even grammatical (for example, sentences (3a,b) and (4a,b) are not fully grammatical). A prominent explanation says that a long-distance dependency between two elements in a sentence is difficult, and often impossible, in the presence of an intervener (for example *speaker* in (1a)). An *intervener* is an element that is *similar* to the two elements that are in a long-distance relation, and structurally intervenes between the two, blocking the relation (Rizzi, 2004). Detailed investigations have shown that long-distance dependencies exhibit gradations of acceptability depending on properties of the intervener (Rizzi, 2004; Grillo, 2008; Friedmann et al., 2009). Franck et al. (2015); Villata and Franck (2016) concentrate on those features that are properties of words: *lexical restriction, number* and *animacy*. This is interesting for us as these are lexical features and therefore they can potentially be captured by word embeddings.

Object Relatives

(1a) Julie smiles to the **student** that the **speaker** has been seriously annoying from the beginning.
(1b) Julie smiles to the **students** that the **speaker** has been seriously annoying from the beginning.
(2a) Julie points out to the **student** that the **speaker** has been yawning frequently from the beginning.
(2b) Julie points out to the **students** that the **speaker** has been yawning frequently from the beginning.

Weak islands

(3a) **Which class** do you wonder **which student** liked?
(3b) **Which professor** do you wonder **which student** liked?
(4a) **What** do you wonder **who** liked?
(4b) **Who** do you wonder **who** liked?

Figure 1: The linguistic constructions and experimental materials, English version. (1) object relatives; (2) completives (experimental control, no long distance dependency), (a) number match, (b) number mismatch. (3) Lexically specified; (4) lexically bare; (a) animacy mismatch; (b) animacy match.

All else being equal, in complex question environments (weak islands, such as those shown in (3) and (4)), long-distance dependency involving a lexically restricted *wh*-phrase (*which class* or *which student*) is more acceptable than extraction of a bare *wh*-element (*who* or *what*), which is not very good.

Experiments on relative clauses also show that the morpho-syntactic feature *number* triggers intervention effects (Belletti et al., 2012; Bentea, 2016). So, for example, the sentence in (1b) is reported to be easier than the sentence in (1a), because the words *students* and *speaker* do not match in number. Completive sentences like those in (2), on the other hand, do not show any difference

between (2a), where number matches, and (2b), where number does not match, as no long-distance dependency is at stake.

The status of a lexical-semantic feature such as *animacy* remains more controversial, but some recent studies show a clear effect of animacy as an intervention feature in *wh*-islands (Franck et al., 2015; Villata and Franck, 2016). So for example, (3a) is easier than (3b) and (4a) is easier than (4b) because the two *wh*-phrases do not match in animacy.

3 The human experiments

The psycholinguistic experiments that collected the experimental measures reflecting the acceptability or reading times of a sentence are described in Franck et al. (2015) and Villata and Franck (2016) and they are the same as those discussed in Merlo and Ackermann (2018). The initial experiments were done in French.[1] Figure 1 shows the English version of the kind of sentences that are used as stimuli in the experiments. The object relative clause experiment collected on-line reading times, manipulating the number (singular or plural) of the object of the relative clause as the intervening feature and the construction, with or without long-distance dependency. A speed-up effect in number mismatch configurations (plural object) was found in object relative clauses. The weak islands experiment collected off-line acceptability judgments, manipulating animacy and lexical restriction of the intervener. A clear effect of animacy match for lexically restricted phrases (less acceptable) and less so for bare *wh*-phrases was found.

Recall that, in Merlo and Ackermann (2018), it was found that similarity scores calculated on these experimental items using word embeddings do not correlate with experimental results. This lead to the conclusion that word embeddings do not encode relevant information related to the important notion of intervener. In the next two sections, we present our extensions to these results.

4 Divergent vectors for French

Artetxe et al. (2018) propose a post-processing vector transformation technique based on eigendecomposition that corresponds to calculating first, second, nth-order similarities. The basic intuition is that, for example, a second-order similarity is a similarity matrix of similarities. Instead of changing the similarity matrix, the word embeddings themselves are transformed, calculating first, second, nth-order similarities directly. These similarities are based on the tuning of a single parameter α, the power of the matrix, to increase or decrease the similarity order. For example α= 0 is first-order similarity, the similarity of two given words, α= 0.5 is second-order similarity, the similarity of the context of two given words. Values of α can vary both positively and negatively. Intuitively, negative values are similarities of two given words as first, second, n-order contexts of other words. Artexte and colleagues argue that different-order similarities are related to different levels of linguistic representations, and certain values of the parameter move the vectors in a space where similarities are more syntactic (as in *sing, singing*), while other values of the parameter move the vectors in a space that is more semantic (as in *sing, chant*). They also distinguish a notion of similarity as analogy, such as the one exhibited by words like *car* and *automobile*, and relatedness, such as in *car* and *road*. Specifically, they claim that their results show that the notion of similarity represented in vectorial space can be decomposed into a more 'syntactic' notion of similarity and the notion of 'relatedness' of a more semantic flavour. They confirm these claims by better performance of the transformed vectors in different tasks of analogy and relatedness that tap into different notions of similarity.

We apply this eigendecomposition technique to our data, serching for the appropriate values of α, to see if moving the vectors in a region of the space that corresponds better to syntactic similarity yields better correlations between word embeddings similarity scores and experimental results than found in Merlo and Ackermann (2018).

Materials and Method The language we use in this experiment is French. Recall that we want to calculate a correlation at the lexical level, as the notion of intervention is based on lexical properties. So we modify the word embeddings of the lexical items.

List of words We use all the words in the stimuli of the human experiments described above, including the fillers, to create an exact replication of the linguistic environment of the experiments. In

[1] We are very grateful to Sandra Villata and Julie Franck for sharing their stimuli and experimental results with us.

total, we have 388 unique words. We use only the words in the actual experimental stimuli for testing (96 words for object relatives and 128 words for weak islands). These are words shown in bold in Figure 1. For this experiment, in the case of lexically restricted weak islands, we only look at the head word; for example, for the phrase *which professor* we use *professor*.

Vectors For all these words, we extract vectors from preexisting trained vectors. We use Fasttext in its new version (Grave et al., 2018), as it is shown in Artetxe et al. (2018) that these vectors yield best results for their eigendecomposition technique. These publicly available vectors have been obtained on a 5-word window, for 300 resulting dimensions, on Wikipedia data using the skip-gram model described in Bojanowski et al. (2016).[2] In this model, every word is represented as an n-grams of characters, for n between 3 and 6. Each n-gram is represented by a vector and the sum of these vectors forms the vector representing the given word.

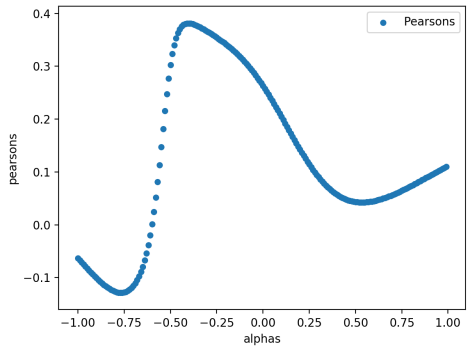
Transformation We apply Artetxe et al. (2018)'s transformation to the word embeddings of the experimental items, for several values of α, in the range -1 to 1.

Calculation of results We calculate correlations with experimental results. Specifically, we first calculate the cosine similarity between the transformed word embeddings of the head of the long-distance dependency and the intervening element (the words in bold in Figure 1). Then, we calculate a correlation between the obtained similarity scores and the experimental measures (mean acceptability judgments for weak islands and reaction times at the position of the verb for relative clauses).

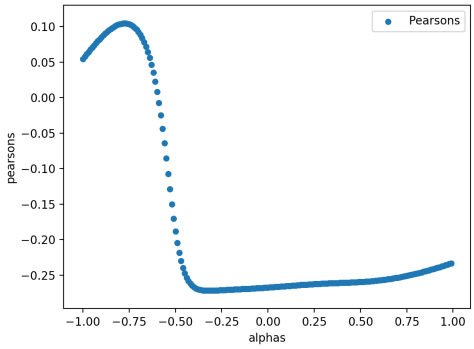
Results and discussion Results for direct correlations are shown in Figure 2. The two panels show how the values of the Pearson correlations between the similarity scores of the transformed word embeddings and the experimental results vary with different values of α. For example, the panel 2a indicates that the maximum correlation between the similarity scores of the word embeddings of the experimental items and the experimental scores are reached at $-0.50 > \alpha > -0.25$.

Overall, for values of α close to 0, the value of α indicating a direct correlation between the words

[2]The French vectors are to be found here https://fasttext.cc/docs/en/crawl-vectors.html.



(a) Object relatives, number.



(b) Weak islands, animacy.

Figure 2: Pearson correlation (range of value of α for transformation: -1 to 1).

of interest, the correlation is very weak, positively for object relatives and the number feature and negatively for weak islands and the animacy feature.

The panel (a) shows how the correlations with the intervention effect of *number* vary with different values of α. It shows a weak correlation reaching 0.4 for values of α between -0.5 and -0.25. This only weakly confirms Merlo and Ackermann (2018)'s results, showing there is some syntactic signal, although not sufficient to explain the experimental results. The panel (b) shows the correlations with the experiment testing the intervention effects of *animacy*. Here the correlation is even weaker, despite the transformation, as positive Pearson values never exceed 0.10 (and the strongest correlation is a negative -0.25), confirming Merlo and Ackermann (2018)'s results, even in this more propitious set up.

An interesting linguistic observation that is quite clear from the patterns of α, is that both the feature *number*, more intuitively syntactic, and the feature *animacy*, whose status is ambiguous

between syntax and semantics, find best correlations with human judgments in similar values of α (very small negatives). This indicates that animacy plays the role of a syntactic feature, in this context, analogously to what has been found in the human experiments.

Another striking feature of the results is the steep curve, in both panels, showing big changes as we move from the words of interest to words in the context. Finally, in both cases the best, but still weak, correlations are in a window of negative values. That is, we find the best correlation when the words in question are treated as context for other (unknown) words. In other words, human experimental results have the best correlation with (unknown) words whose paradigmatic context (the second-order context) is defined by the word embeddings. We can conjecture that this indicates that the words in the stimuli sentence (the words in bold in Figure 1), taken as a second-order context, define a lexical semantic field to which the experimental measures are sensitive.[3]

5 Prediction in weak islands and object relatives in French and English

While the previous experiments show that even transformed word embeddings do not encode fine-grained quantitative psycholinguistic measures, it is still possible that sentence embeddings can predict the coarser distinctions of qualitative acceptability judgments. Moving away from seeking direct correlations with experimental results to a prediction task that simply models acceptability judgments presents the added advantage that it is easily extended to new languages. The acceptability judgments in these constructions are easily established; in fact, in our case, they match exactly the judgments in French.

5.1 Materials and Method

Sentences All French weak islands experimental stimuli used in the previous experiments were translated into English. The translations were literal. They were performed by a bilingual near-native speaker of English. They correspond to sentences established in the literature on weak islands in English.[4]

Sentence embeddings We calculate sentence embeddings for all the sentences. Here, we follow the logic of probing tasks (Conneau et al., 2018). In this set up, single sentence embeddings are classified according to a single grammatical phenomenon. This technique allows the researchers to reach clear and conclusive insights into what information is encoded in the embeddings in a set-up that is agnostic of the architecture that has produced them. Our problem also meets other criteria that have been advocated for probing tasks, some formal and some more subjective. Namely, the domain of linguistic locality of the phenomenon we want to study is a single sentence (as opposed to multi-sentence tasks), and all the sentences are carefully controlled and matched to eliminate sentence length effects, for example.[5] Finally, we agree that probing tasks should address a set of linguistically interesting phenomena. From this standpoint, intervention effects in long-distance dependencies meet the requirement, as one of the core data paradigms definitional of human languages.

We use bag of vectors sentence embeddings for several reasons. First, our testing sentences are carefully constructed minimal pairs where the difference in grammaticality hinges on one lexical difference, which then has in turn syntactic repercussions. By using bag of vectors, we remain as close as possible to the lexical setting of the theoretical definition of intervention, We also differ only minimally from the previous experiment. Second, from a more practical standpoint, bag of vectors have been shown in general, and in probing tasks in particular, to have good performance. Notice also that the choice of not using more context-aware vector representations is voluntary. We want to test the predictive ability of a direct encoding of the linguistic notion of intervener, which is a lexical, non-contextualised notion.

We use BoVfastText,[6] which derives sentence representations by averaging the fastText embeddings of the words they contain.

Classifiers We use a multi-layer perceptron, with four outputs, and two hidden layers of 50 and

[3] In future work, this conjecture could be verified by retrieving these unknown words and seeing if a direct correlation is confirmed.

[4] See supplementary materials for all the English datasets discussed in the paper.

[5] Differently from other probing tasks, we work here on a relatively small dataset, but in principle the sentences follow a specific structure that could be easily automated if more data needed to be tested. But the small amount of testing data already gives us an effect.

[6] https://fasttext.cc/docs/en/english-vectors.html

Label	French	English
BareA	0.151	0.485
BareI	0.909	0.156
LexA	0.788	0.273
LexI	0.303	0.091

(a) Weak islands (Bare= bare wh, Lex= lexically specified, A= animate, I= inanimate).

Label	French	English
ORCsing	0.250	0.417
ORCplur	0.125	0.375
CMPsing	0.291	0.291
CMPplur	0.500	0.292

(b) Object Relatives (CMP= completive).

Table 1: Percent accuracy predictions.

30 dimensions.[7] We run the training and testing in an n-fold cross-validation regime (n = 33 for weak islands and 24 for object relatives), where each quadruple of examples is used for testing.

Dependent variable We use accuracy as a measure of how much the information in the input embeddings supports the discrimination of the four sentence types in a categorical classifier. This measure is relevant under the assumption that the more acceptable sentences are more easily identified (discriminated from other classes) than less acceptable ones, because acceptable sentences better fit to the grammar. Less acceptable sentences do not fit or even do not belong to the grammar, and as such their classes are more easily confusable, given that the complement of a grammar does not necessarily have distinguishing structured characteristics. So we expect to see higher classification accuracy as the acceptability of the sentence increases.

5.2 Predictions and results

The accuracy prediction task corresponds to the structure of the human experiment.

In the materials on weak islands, we have four sentence types (BareA, BareI, LexA, LexI), shorthand for the four cases in which the stimuli had an animate (A) or inanimate (I) intervener and where the long-distance dependency was lexically specified (Lex) or bare (Bare).

Let Acc() be the accuracy of the prediction. Recall that animacy is the property that leads to intervention, and expected degraded performance, so we expect Acc(LexA) $<$ Acc(LexI) and Acc(BareA) $<$ Acc(BareI). Since we know that lexical specification improves acceptability, we expect Acc(LexA) $>$ Acc(BareA) and Acc(LexI) $>$ Acc(BareI).

We can see the results in Table 1, subtable 1a. For French, the prediction on the effect of animacy in the lexically specified case is confirmed, but the others are not. Furthermore, if we calculate the total interaction, we see that lexical specification makes these sentences easier to a greater extent than animacy makes them hard.[8] This result corroborates effects found in human experiments, which found a stronger effect of animacy than lexical specification (Franck et al., 2015; Villata and Franck, 2016). For English, we find that, given the same predictions as above, the prediction for the effect of animacy is confirmed both in bare *wh*-phrases and in lexicalised *wh*-phrases, but the others are not. A total interaction does not confirm the dominant effect of animacy, unlike French. [9]

For object relative clauses, it is a match in number of the relative head and the subject of the relative clause (both singular) that is expected to cause difficulty, compared to a mismatch. Object relative clauses are also compared to completives, where no differences should be found between the two items. So we expect, Acc(ORCsing) $<$ Acc(ORCplur) and Acc(CMPsing) $=$ Acc(CMPplur). Also, Acc(ORCsing) $<$ Acc(CMPsing) and Acc(ORCplur) $=$ Acc(CMPplur). We can see the results in Table 1, subtable 1b. For French, none of the predictions is confirmed, while for English the only confirmed prediction says that *number*, whether singular or plural should be roughly similar in completives, the control case.

More results of this same nature and reaching similar conclusions can be found in the Appendix, for both weak islands and object relative clauses in both English and French. The appendix shows results obtained with a Naive Bayes classifier, thereby also demonstrating that the negative effect is not due to the choice of classifier.

[7] Two larger hidden layers of 200 and 100 dimensions also gave similar results

[8] (Acc(BareA) $-$ Acc(BareI)) $-$ (Acc(LexA) $-$ Acc(LexI)), we find $(0.909 - 0.788) - (0.151 - 0.303) = 0.121 + 0.152 > 0$.

[9] (Acc(BareA) $-$ Acc(BareI)) $-$ (Acc(LexA) $-$ Acc(LexI)), we find $(0.156 - 0.273) - (0.485 - 0.091) == -0.117 - 0.394 < 0$.

6 Discussion and conclusions

These results prompt both scientific and methodological considerations. Scientifically, our results lead us to conclude that while current word embeddings encode some notion of similarity, as shown by many experiments on analogical tasks and textual and lexical similarity, they do not, however, encode the notion of similarity that has been shown to be at work in many human experiments and to be definitional in long-distance dependencies.

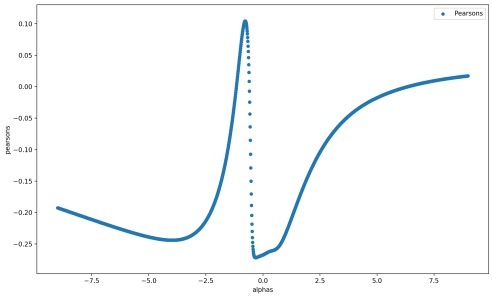
If our conclusions above are correct, our lack of replication of some of the theoretical predictions and human experiments adds one more discordant element to the complex debate of whether a narrow or broad definition of intervention best explains human judgments and linguistic facts, as discussed by Villata and Franck. The narrow notion of intervention is grammar-based, explains ungrammaticality, for example weak islands, and claims that only morpho-syntactic features are relevant to define intervention (Rizzi, 2004). So, the fact that word embeddings —usage-based representations of the lexical semantics of words— do not correlate with a grammar-based notion of similarity is to be expected, but the fact that object relative clauses where found to exhibit animacy effects in human experiments is not expected.

A broader notion of intervention is defined by cue-based memory based models: these are human sentence processing models that explain difficulty of otherwise grammatical sentences, such as object relatives (Van Dyke and McElree, 2006). In this framework, similarity can take any feature type into account and intervention is a kind of interference at retrieval in memory. This broader approach explains the experimental findings, but would have expected a correlation of word embeddings, which are fundamentally a semantic encoding of the word, with the experimental effects.

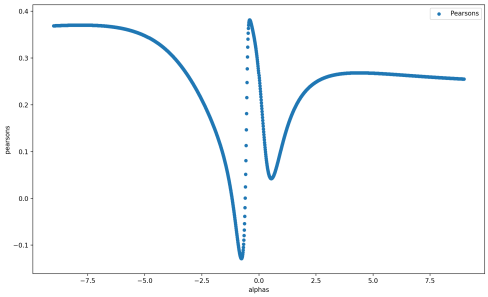
Methodologically, we might wonder about the sources of the fluctuation of results, both for long-distance dependence as reported here, and subject verb agreement as reported in other works mentioned in the introduction. Two explanations are possible: the methods are not sound, the fluctuations are to be expected. I very briefly explore both.

Consider the transformations we have applied in section 4. Word embeddings are a merger of many kinds of information and applying post-processing transformations has been argued to tease apart syntactic aspects of the encoding of the notion of similarity from semantic aspects. The notion of 'syntactic' and 'semantic' similarity used in previous work is itself vague and does not refer to any linguistic phenomenon that current linguistic theory (syntactic or semantic) would identify as belonging exclusively to one or other of these levels of representation. Trying to investigate RNN by claiming that certain constructions reflect syntax while other reflect semantics is therefore an ill-defined endeavour (Artetxe et al., 2018). All constructions have a syntax and a semantics and the investigation of what RNN learn can only be done by correlating the predictions of the syntactic or semantic *theory* involved in the construction. To prove this point further, consider the plots in Figure 3. Here, as an abstract exercise, the αs have been varied on a much larger range of values than the more limited range of values reported in section 4. The interval of values in section 4 was chosen because it corresponds to previous proposals and because it is more easily interpretable. As it can be seen here, instead, the curves have a dramatic range of correlation values that do not seem to have any correspondence to anything we know about language. We would conclude here that Blackbox investigations must be driven by theory, or at least by precise expectations grounded in well-established linguistic facts, to become interpretable.

On the other hand, we are also at the beginning of this trend of Blackbox investigations. We submit here that these fluctuations are an effect known as the Proteus effect, fluctuations due to the fact that we are in the early stages of this promising avenue of research, due to the fast publishing rate and to the small size of the studies. The Proteus phenomenon—a term coined by Ioannidis and Trikalinos (2005)—describes the effect of rapidly alternating opposite research claims and extremely opposite refutations, particularly during the early accumulation of data. Meta-research and simulations show that first publication of results have a considerably higher chance of being inflated (Ioannidis, 2008), and that small studies have a higher chance of being false positives (Bertamini and Munafò, 2012). We submit, then, that the contradictory results are inevitable incongruities that will be resolved as more studies, large and small, accumulate on these same topics.



(a) Weak islands



(b) Object relatives

Figure 3: Pairwise correlation (range of value of alpha for transformation: -9 to 9).

7 Conclusions

In this work, we have extended previous work on long-distance dependencies applying new vector transformation techniques, extending the investigation to sentence embeddings and to new languages. We confirm previous negative results: word embeddings and sentence embeddings, the representations produced by RNNs, do not unequivocally encode fine-grained linguistic properties of long-distance dependencies. Future work, among many other avenues for extension, will investigate in more detail the limits of vector transformation techniques and extend the work to different vectorial encodings, to more constructions and to new languages.

Acknowledgments

We thank Julie Franck and Sandra Villata for allowing the use of their French stimuli and their translation in English and Mikel Artexte for useful interactions. All remaining errors and opinions are our own.

References

Mikel Artetxe, Gorka Labaka, Iñigo Lopez-Gazpio, and Eneko Agirre. 2018. Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pages 282–291.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transaction of the ACL*, 7:49–72.

Adriana Belletti, Naama Friedmann, Dominique Brunato, and Luigi Rizzi. 2012. Does gender make a difference? Comparing the effect of gender on children's comprehension of relative clauses in Hebrew and Italian. *Lingua*, 122(10):1053–1069.

Anamaria Bentea. 2016. *Intervention effects in language acquisition: the comprehension of A-bar dependencies in French and Romanian*. Ph.D. thesis, University of Geneva.

Marco Bertamini and Marcus R. Munafò. 2012. Bite-size science and its undesired side effects. *Perspectives on Psychological Science*, 7:67–71.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

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 (COLING'18)*, pages 133–144. Association for Computational Linguistics.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. 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. Association for Computational Linguistics.

Julie Franck, Saveria Colonna, and Luigi Rizzi. 2015. Task-dependency and structure dependency in number interference effects in sentence comprehension. *Frontiers in Psychology*, 6.

Naama Friedmann, Adriana Belletti, and Luigi Rizzi. 2009. Relativized relatives: Types of intervention in the acquisition of A-bar dependencies. *Lingua*, 119(1):67 – 88.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Nino Grillo. 2008. *Generalized minimality: Syntactic underspecification in Broca's aphasia*. Ph.D. thesis, University of Utrecht.

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

John Ioannidis. 2008. Why most discovered true associations are inflated. *Epidemiology*, 19(5):640–648. Doi:10.1097/EDE.0b013e31818131e7.

John Ioannidis and Thomas A. Trikalinos. 2005. Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomised trials. *Journal of Clinical Epidemiology*, pages 543–549.

Adhiguna Kuncoro, Chris Dyer, John Hale, and Phil Blunsom. 2018. The perils of natural behaviour tests for unnatural models: the case of number agreement. Poster at Learning Language in Humans and Machines (L2HM 2018).

Yair Lakretz, Germán Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. *CoRR*, abs/1903.07435.

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

Paola Merlo and Francesco Ackermann. 2018. Vectorial semantic spaces do not encode human judgments of intervention similarity. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pages 392–401.

Luigi Rizzi. 2004. Locality and left periphery. In Adriana Belletti, editor, *The cartography of syntactic structures*, number 3 in Structures and beyond, pages 223–251. Oxford University Press, New York.

Julie A. Van Dyke and Brian McElree. 2006. Retrieval interference in sentence comprehension. *Journal of memory and language*, 55(2):157–166.

Sandra Villata and Julie Franck. 2016. Semantic similarity effects on weak islands acceptability. In *41st Incontro di Grammatica Generativa Conference*, Perugia, Italy. Https://archive-ouverte.unige.ch/unige:82418.

Label	Precision	Recall	F-ratio
BareA	0.257	0.625	0.365
BareI	0.196	0.350	0.252
LexA	0.454	0.062	0.110
LexI	0.238	0.065	0.102

(a) Weak Islands, English.

Label	Precision	Recall	F-ratio
BareA	0.707	0.854	0.773
BareI	0.410	0.854	0.555
LexA	0.181	0.054	0.084
LexI	0.340	0.205	0.256

(b) Weak Island, French

Table 2: Results for weak islands (Bare= bare wh, Lex= lexically specified, A= animate, I= inanimate).

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. Association for Computational Linguistics.

Appendix A: Comparing results across classifiers

As further demonstration of the fluctuating nature of the current results, we report here the same classification experiment reported above, but with a Naive Bayes classifier, and applied to both weak islands and object relative clauses. The classification results averaged over ten trials (same cross-validation settings as above) are shown in Table 2 and Table 3.

Recall that animacy is the property that leads to intervention in weak islands, and expected degraded performance, so we expect that Acc(LexA) $<$ Acc(LexI) and Acc(BareA) $<$ Acc(BareI). Furthermore, Acc(LexA) $>$ Acc(BareA) and Acc(LexI) $>$ Acc(BareI).

For object relative clauses, it is a match in number of the relative head and the subject of the relative clause (both singular) that is expected to cause difficulty, compared to a mismatch. Object relative clauses are also compared to completives, where no differences should be found between the two items. So we expect, Acc(ORCsg) $<$ Acc(ORCpl) and Acc(CMPsg) $=$ Acc(CMPpl). Also, Acc(ORCsg) $<$ Acc(CMPsg) and Acc(ORCpl) $=$ Acc(CMPpl).

Weak islands do not conform to expectations: For English, we can see that neither

Label	Precision	Recall	F-ratio
CMPsg	0.02	0.08	0.032
CMPpl	0.17	0.125	0.14
ORCsg	0.35	0.2	0.25
ORCpl	0.16	0.18	0.17

(a) Object relatives, English

Label	Precision	Recall	F-ratio
CMPsg	0.18	0.25	0.209
CMPpl	0.23	0.29	0.256
ORCsg	0.105	0.10	0.102
ORCpl	0.21	0.08	0.116

(b) Object relatives, French

Table 3: Results for object relatives.

Acc(LexA) $<$ Acc(LexI) nor Acc(BareA) $<$ Acc(BareI) are confirmed. Moreover, Acc(LexA) $>$ Acc(BareA) is not confirmed and neither is Acc(LexI) $>$ Acc(BareI). For French, we can see that Acc(LexA) $<$ Acc(LexI) is confirmed, but Acc(BareA) $<$ Acc(BareI) is not. Moreover, Acc(LexA) $>$ Acc(BareA) is not confirmed and neither is Acc(LexI) $>$ Acc(BareI).

For English Acc(ORCsg) $<$ Acc(ORCpl) is not confirmed and Acc(CMPsg) $=$ Acc(CMPpl) is also not confirmed as the difference is quite significant. Also, Acc(ORCsg) $<$ Acc(CMPsg) is not confirmed but Acc(ORCpl) could be considered not very different from Acc(CMPpl) .

For French, Acc(ORCsg) $<$ Acc(ORCpl) is not confirmed, but Acc(CMPsg) and Acc(CMPpl) are not very different Also, Acc(ORCsg) $<$ Acc(CMPsg) is confirmed but Acc(ORCpl) is smaller than Acc(CMPpl).

Appendix B: English sentences

Weak Islands

1. What model do you wonder what man painted?
2. Who do you wonder who painted?
3. What do you wonder who painted?
4. What landscape do you wonder what man painted?
5. What book do you wonder what student has forgotten?
6. What do you wonder who has forgotten?
7. What friend do you wonder what student has forgotten?
8. Who do you wonder who has forgotten?
9. Who do you wonder who has eaten?
10. What rooster do you wonder what fox has eaten?
11. What do you wonder who has eaten?
12. What cheese do you wonder what fox has eaten?
13. What trousers do you wonder what tailor has looked for?
14. What do wonder who has looked for?
15. Who do you wonder who has looked for?
16. What customer do you wonder what tailor has looked for?
17. Who do you wonder who was looking at?
18. What producer do you wonder what actor was looking at?
19. What do you wonder who was looking at?
20. What do you wonder what actor was looking at?
21. What do you wonder who has brought?
22. What bag do you wonder what traveller has brought?
23. What friend do you wonder what traveler has brought?
24. Who don ou wonder who has brought?
25. What professor do you wonder what student has appreciated?
26. Who do you wonder who has appreciated?
27. What do you wonder who has appreciated?
28. What course do you wonder what student has appreciated?
29. What exam do you wonder what intern has feared?
30. What do you wonder who has feared?

31. What doctor do you wonder what intern has feared?
32. Who do you wonder who has feared?
33. Who do you wonder who has heard?
34. What keeper do you wonder what animal has heard?
35. What noise do you wonder what animal has heard?
36. What do you wonder who has heard?
37. What number do you wonder what baby-sitter has kept?
38. What do you wonder who has kept?
39. Who do you wonder who has kept?
40. What baby do you wonder what baby-sitter has kept?
41. Who do you wonder who has regretted?
42. What colleague do you wonder what director has regretted?
43. What advice do you wonder what counsellor has regretted?
44. What do you wonder who has regretted?
45. What speech do you wonder what guest has listened to?
46. What do you wonder who has listened to?
47. What speaker do you wonder what guest has listened to?
48. Who do you wonder who has listened to?
49. Who do you wonder who has taken pictures of?
50. What model do you wonder what artist has taken pictures of?
51. What painting do you wonder what artist has taken pictures of?
52. What do you wonder who has taken pictures of?
53. What hat do you wonder what designer has chosen?
54. What do you wonder who has chosen?
55. What model do you wonder what designer has chosen?
56. Who do you wonder who has chosen?
57. What customer do you wonder what employee has been waiting for?
58. Who do you wonder who has been waiting for?
59. What do you wonder who has been waiting for?
60. What salary do you wonder what employee has been waiting for?
61. What do you wonder who has appreciated?
62. What gift do you wonder what winner has appreciated?
63. What athlete do you wonder what winner has appreciated?
64. Who don ou wonder who has appreciated?
65. What athlete do you wonder what winner has appreciated?
66. Who do you wonder who has appreciated?
67. What do you wonder who has appreciated?
68. What gift do you wonder what winner has appreciated?
69. What hero do you know what veteran had met?
70. Who do you know who has met?
71. What day ou know who has met?
72. What challenge do you know what veteran has met?
73. What necklace do you know what student has lost?
74. What do you know who has lost?
75. What friend do you know what student has lost?
76. Who do you know who has lost?
77. What actor do you know what viewer loved?

78. Who do you know who loved?
79. What movie do you know what viewer loved?
80. What do you know who loved?
81. What do you know who carried?
82. What suit do you know what singer carried?
83. What fan do you know what singer carried?
84. Who do you know who carried?
85. What bore do you know what pedestrian has found?
86. Who do you know who has found?
87. What do you know who has found?
88. What wallet do you know what pedestrian has found?
89. What do you know who has found?
90. Who do you know who has found?
91. What treasure do you know what child has found?
92. What friend do you know what child has found?
93. Who do you know who has abandoned?
94. What child do you know what man has abandoned?
95. What do you know who has abandoned?
96. What apartment do you know what man has abandoned?
97. What do you know who has filmed?
98. What documentary do you know what cameraman has filmed?
99. Who do you know who has filmed?
100. What actor do you know what cameraman has filmed?
101. What attacker dont you know what man has defeated?
102. Who dont you know who has defeated?
103. What cancer dont you know what man has defeated?
104. What dont you know who has defeated?
105. What dont you know who has kidnapped?
106. What evidence dont you know what kidnapper has concealed?
107. Who dont you know who has kidnapped?
108. What orphan dont you know what kidnapper has concealed?
109. Who dont you know who has left?
110. What friend dont you know what researcher has left?
111. What country dont you know what researcher has left?
112. What dont you know who has left?
113. What dont you know who has followed?
114. What studies dont you know what doctoral student has followed?
115. What intern dont you know what doctoral student has followed?
116. Who dont you know who has followed?
117. Who dont you know who has run over?
118. What pedestrian dont you know what driver has run over?
119. What bicycle dont you know what driver has run over?
120. What dont you know who has run over?
121. What difficulties dont you know what apprentice has met?
122. What dont you know who has met?
123. Who dont you know who has met?
124. What instructor dont you know what apprentice has met?
125. What criminal dont you know what lawyer denounced?
126. Who dont you know who denounced?
127. What dont you know who denounced?

128. What abuse dont you know what lawyer denounced?

129. What dont you know who decorated?

130. What banner dont you know what general decorated?

131. Who dont you know who decorated?

132. What lieutenant dont you know what general decorated?

Object Relative Clauses

1. Julia points out to the student that the speaker has been yawning frequently from the beginning.

2. Paul explains to the voter that the politician has been clearly lying since the elections.

3. Sebastian reveals to the patient that the tranquilliser has been acting progressively for a year.

4. Jerome points out to the prisoner that the warden sometimes comes into the courtyard.

5. Charles explains to the victim that the treatment is starting slowly but surely.

6. Benjamin reminds the teen-ager that the educator has been drinking frequently for a few years.

7. Bernard reminds the gamblers that the casino is closing unfortunately very soon.

8. Laura says to the shepherds that the sheep is bleeting stupidly after the shearing.

9. Peter announces to the candidates that the jury will deliberate firmly after the audition.

10. Mark repeats to the people that the unhappiness continues inevitably after the tragedy.

11. Claire reminds the workers that the fireplace has been smoking a lot since the works.

12. Patricia says to the customers that the hat is very pleasing because of the feathers.

13. Fred smiles to the child that the priest has been blessing happily after each service.

14. Lise speaks to the woman whose weight the diet is reducing surprisingly easily.

15. Giles speaks to the worker that the effort has been tiring inevitably with time.

16. Jack thinks of the owner that the stress has been aging prematurely despite the anti-anxiety medications.

17. Patrick thinks of the family that the holidays have been reuniting every year for ten days.

18. Louise smiles to the girl that the witch frightens on purpose for Halloween.

19. Aude is liked by the athletes that the massage relaxes always after the training.

20. Luke thinks of the girls that the seducer has been adressing assiduously for an hour.

21. Anne speaks to the actors that the audience has been applauding frantically after each show.

22. Joan speaks to the neighbours that the excursion has rarely enthused at the end of the year.

23. Joan calls the lawyers that the disappointment embitters inevitably after the trial.

24. Roland smiles to the offenders that the policeman has been investigating secretly for a month.

25. Julia smiles to the students that the speaker has been putting to sleep seriously from the beginning.

26. Paul thinks of the voters that the politician has been frankly disappointing since the elections.

27. Sebastian smiles to the patients that the tranquilliser has been weakening progressively for a year.

28. Jerome talks to the prisoners that the warden sometimes lets out into the courtyard.

29. Charles smiles to the victims that the treatment is curing slowly but surely.

30. Benjamin thinks of the teen-agers that the educator has been beating often for a few years.

31. Bernard reminds the gambler that the casino ruins unfortunately very fast.

32. Laura smiles to the shepherd that the sheep is following stupidly after the shearing.

33. Peter calls the candidate that the jury will firmly wait for after the audition.

34. Mark thinks of the people that the unhappiness unites inevitably after the tragedy.

35. Claire attracts the worker that the fireplace has blackened a lot since the works.

36. Patricia talks to the customer that the hat makes tall because of the feathers.

37. Fred tells the children that the priest has been leaving happily after each service.

38. Lise promises the women that the diet can be managed surprisingly easily.

39. Giles reminds the workers that the effort will double inevitably with time.

40. Jack explains the owners that stress arrives prematurely despite the anti-anxiety medications.

41. Patrick repeats to the families that the holidays last every year for ten days.

42. Louise repeats to the girls that the witch makes grimaces on purpose for Halloween.

43. Aude repeats to the athlete that the massage always begins after the training.

44. Luke tells the girl that the seducer has been chatting untiringly for an hour.

45. Anne reminds the actor that the audience has been laughing frantically after each show.

46. Joan reminds the neighbour that the excursion has rarely failed at the end of the year.

47. Joan reminds the lawyer that the disappointment remains inevitably after the trial.

48. Roland points out to the offender that the policeman has been intervening secretly for a month.

49. Julia points out to the students that the speaker has been yawning frequently from the beginning.

50. Paul explains to the voters that the politician has been clearly lying since the elections.

51. Sebastian reveals to the patients that the tranquilliser has been acting progressively for a year.

52. Jerome points out to the prisoners that the warden sometimes comes into the courtyard.

53. Charles explains to the victims that the treatment is starting slowly but surely.

54. Benjamin reminds the teen-agers that the educator has been drinking frequently for a few years.

55. Bernard reminds the gambler that the casino is closing unfortunately very soon.

56. Laura says to the shepherd that the sheep is bleeting stupidly after the shearing.

57. Peter announces to the candidate that the jury will deliberate firmly after the audition.

58. Mark repeats to the population that the unhappiness continues inevitably after the tragedy.

59. Claire reminds the worker that the fireplace has been smoking a lot since the works.

60. Patricia says to the customer that the hat is very pleasing because of the feathers.

61. Fred smiles to the children that the priest has been blessing happily after each service.

62. Lise speaks to the women whose weight the diet is reducing surprisingly easily.

63. Giles speaks to the workers that the effort has been tiring inevitably with time.

64. Jack thinks of the owners that the stress has been aging prematurely despite the anti-anxiety medications.

65. Patrick thinks of the families that the holidays have been reuniting every year for ten days.

66. Louise smiles to the girls that the witch frightens on purpose for Halloween.

67. Aude is liked by the athlete that the massage relaxes always after the training.

68. Luke thinks of the girl that the seducer has been adressing assiduously for an hour.

69. Anne speaks to the actor that the audience has been applauding frantically after each show.

70. Joan speaks to the neighbour that the excursion has rarely enthused at the end of the year.

71. Joan calls the lawyer that the disappointment embitters inevitably after the trial.

72. Roland smiles to the offender that the policeman has been investigating secretly for a month.

73. Julia smiles to the student that the speaker has been putting to sleep seriously from the beginning.

74. Paul thinks of the voter that the politician has been frankly disappointing since the elections.

75. Sebastian smiles to the patient that the tranquilliser has been weakening progressively for a year.

76. Jerome talks to the prisoner that the warden sometimes lets out into the courtyard.

77. Charles smiles to the victim that the treatment is curing slowly but surely.

78. Benjamin thinks of the teen-ager that the educator has been beating often for a few years.

79. Bernard reminds the gamblers that the casino ruins unfortunately very fast.

80. Laura smiles to the shepherds that the sheep is following stupidly after the shearing.

81. Peter calls the candidates that the jury will firmly wait for after the audition.

82. Mark thinks of the people that the unhappiness unites inevitably after the tragedy.

83. Claire attracts the workers that the fireplace has blackened a lot since the works.

84. Patricia talks to the customers that the hat makes tall because of the feathers.

85. Fred tells the child that the priest has been leaving happily after each service.

86. Lise promises the woman that the diet can be managed surprisingly easily.

87. Giles reminds the worker that the effort will double inevitably with time.

88. Jack explains the owner that stress arrives prematurely despite the anti-anxiety medications.

89. Patrick repeats to the family that the holidays last every year for ten days.

90. Louise repeats to the girl that the witch makes grimaces on purpose for Halloween.

91. Aude repeats to the athletes that the massage always begins after the training.

92. Luke tells the girls that the seducer has been chatting untiringly for an hour.

93. Anne reminds the actors that the audience has been laughing frantically after each show.

94. Joan reminds the neighbours that the excursion has rarely failed at the end of the year.

95. Joan reminds the lawyers that the disappointment remainsinevitably after the trial.

96. Roland points out to the offenders that the policeman has been intervening secretly for a month.