Vectorial semantic spaces do not encode human judgments of intervention similarity

Paola Merlo and **Francesco Ackermann**
Department of Linguistics
University of Geneva
5 Rue de Candolle, CH-1211 Genève 4
paola.merlo@unige.ch, francesco.ackermann@unige.ch

Abstract

Despite their practical success and impressive performances, neural-network-based and distributed semantics techniques have often been criticized as they remain fundamentally opaque and difficult to interpret. In a vein similar to recent pieces of work investigating the linguistic abilities of these representations, we study another core, defining property of language: the property of long-distance dependencies. Human languages exhibit the ability to interpret discontinuous elements distant from each other in the string as if they were adjacent. This ability is blocked if a similar, but extraneous, element intervenes between the discontinuous components. We present results that show, under exhaustive and precise conditions, that one kind of word embeddings and the similarity spaces they define do not encode the properties of intervention similarity in long-distance dependencies, and that therefore they fail to represent this core linguistic notion.

1 Introduction

Despite their practical success and impressive performances, neural-network-based and distributed semantics techniques have often been criticized as they remain fundamentally opaque and difficult to interpret.

To cast light on what linguistic information is learnt and encoded in these representations, several pieces of work have recently studied core properties of language in syntax (Linzen et al., 2016; Bernardy and Lappin, 2017; Gulordava et al., 2018; Linzen and Leonard, 2018; van Schijndel and Linzen, 2018), semantics (Herbelot and Ganesalingam, 2013; Erk, 2016), morphology (Cotterell and Schütze, 2015). In a similar vein, we study another core, defining property of human languages: the property of long-distance dependencies.

Human languages exhibit the ability to interpret discontinuous elements distant from each other in the string as if they were adjacent.[1] Sentence (1a) is a question about the object of the verb *buy*, whose canonical position is shown in angle brackets, thus connecting the first and last element in the sentence.[2] Sentence (2a) is a relative clause where the object of the verb *wash* is also the semantic object of the verb *show*, connecting two distant elements. Sentence (3a) is also a relative clause where the word *étudiant* (student) is the semantic object of the verb *endort* (put to sleep).

(1a) What do you wonder **John** bought <what> ?

(2a) Show me the elephant that **the lion** is washing <the elephant>.

(3a) Jules sourit aux étudiants que l'orateur endort < étudiants> sérieusement depuis le début.

'Jules smiles to the students who the speaker is putting seriously to sleep from the beginning.'

Long-distance dependencies are not all equally acceptable. The precise description of the facts involving long-distance dependencies is complex, and is one of the major topics of research in current linguistic theory, with many competing proposals

[1] To clarify the perhaps confusing terminology: the term long-distance dependencies is a technical term that refers to discontinuous constructions where two elements in the string receive the same interpretation. Long-distance dependency constructions are *wh*-questions, relative clauses, right-node raising, among others (Rimell et al., 2009; Nivre et al., 2010; Merlo, 2015). Not all long-distance are actually long, for example subject-oriented relative clauses, and not all long dependencies are long-distance dependencies, for example, long subject-verb agreement as studied in Linzen et al. (2016); Bernardy and Lappin (2017); Gulordava et al. (2018) is usually not considered a long-distance dependency.

[2] The unpronounced element(s) in the long-distance relation are indicated by < >.

Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL 2018), pages 392–401
Brussels, Belgium, October 31 - November 1, 2018.

(Rizzi, 1990; Gibson, 1998). We will adopt an intuitive and simple explanation, called intervention theory, some aspects of which will be explained in more detail below (Rizzi, 1990, 2004). In a nutshell, a long-distance dependency between two elements in a sentence is difficult or even impossible if a similar element intervenes. For example, sentence (1a) is acceptable while (2a) causes trouble for children (Friedmann et al., 2009) and (3a) triggers agreement errors, because in (1a) there is no sufficiently *similar* intervener (*John* is animate and is not a question word while *what* introduces a question and is not animate), while in (2) and (3) there is (*lion* is animate like *elephant* and *étudiants* (students) is animate like *orateur* (speaker)).

We present results that show, under precise conditions, that one kind of word embeddings and the similarity spaces they define do not encode the notion of intervention similarity involved in long-distance dependencies, but probably only semantic associations.

2 Long-distance phenomena and word embeddings

All languages allow some form of long-distance dependencies under restrictive conditions: for example, (1a) is allowed, but (1b) is not allowed (sentences like (1b) are called weak islands, we keep this terminology),[3] (2a) is hard for children, while (2b) is not, and neither of them is hard for adults, (3a), repeated here as (3b) often triggers agreement mistakes, as shown.

(1b) * What do you wonder **who** bought <what>?

(2b) Show me the elephant that <the elephant> is washing the lion.

(3b) Jules sourit aux étudiants que l'orateur <**étudiants**> endort/*endorment <étudiants> sérieusement depuis le début.

'Jules smiles to the students who the speaker is/*were putting seriously to sleep from the beginning.'

Core to the explanation of these facts is the notion of *intervener*. 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. In our examples, potential interveners are shown in bold.[4]

a. What do you wonder who bought?
b. Which book do you wonder who bought?
c. Which book do you wonder which linguist bought?

Figure 1: Weak islands (< means better). Acceptability judgments: c < b < a.

This explains why (1a) is ok, since there is a potential intervener, but *John* and *what* are not similar, but (1b) is not ok, since there is an intervener, and *who* and *what* are similar, as they are both *wh*-words. Sentence (2a) is hard for children as *the lion* intervenes between the two positions that give meaning to *the elephant*, but sentence (2b) is not, because nothing intervenes. Sentence (3b) triggers agreement mistakes because the intermediate position of *étudiants* intervenes between the word and the verb, causing interference.

Detailed investigations have shown that long-distance dependencies exhibit gradations of acceptability depending on which features are involved (Rizzi, 2004; Grillo, 2008; Friedmann et al., 2009). For example, all other things being equal, in complex question environments (weak islands), we have the gradation of judgments shown in Figure 1, where long-distance dependency involving a lexically restricted *wh*-phrase (*which book* or *which linguist*) is more acceptable than extraction of a bare *wh*-element (*who* or *what*), which is not very good. Experiments on weak islands and relative clauses also show that number triggers intervention effects (Belletti et al., 2012; Bentea, 2016). Thus, results from theoretical linguistics, acquisition and sentence processing point to a definition of intervener based on

[3] As always, * means ungrammatical.

[4] Notice that here and in all the following, intervention is defined structurally and not linearly. Linear intervention that does not structurally hierarchically dominate (technically c-command) does not matter as shown by the contrast **When do you wonder who won?/You wonder who won at five* compared to *When did the uncertainty about who won dissolve?/The uncertainty about who won dissolved at five.* (Rizzi, 2013) Also, intervention can be visible in the string, like in (1) and (2), or understood, as in (3). The intermediate step in relating the two elements of the long-distance dependency in (3) is postulated on theoretical grounds (see for example (Chomsky, 2001), and receives confirmation by participial agreement in languages like French (Kayne, 1989), or the agreement mistakes in the article we use here (Franck et al., 2015). See also Gibson and Warren (2004) for experimental evidence for the role of intermediate steps in long-distance dependencies.

syntactically-relevant features.[5] The status of a lexical-semantic feature such as *animacy* remains more controversial; some results argue in favor of an ameliorative effect (Brandt et al., 2009), some suggest animacy has no effect (Adani, 2012). Some recent studies show a clear effect of animacy as an intervention feature in *wh*-islands (Franck et al., 2015; Villata and Franck, 2016).

We are going to focus on those features for which relevant data is available, and there's reason to think they could be captured in lexical (semantic) vectors because they are properties of words (in contrast to the more discourse-oriented features, such as +Top.) In particular, we focus on *lexical restriction, number* and *animacy* in the definition of intervention similarity.

Sophisticated definition of lexical proximity in feature spaces, called *word embeddings*, have been defined recently in computational linguistics. These embeddings are the vectorial representation of the meaning of a word, defined as the usage of a word in its context (Wittgenstein, 1953 [2001]; Harris, 1954; Firth, 1957). Tasks that confirm this interpretation are association, analogy, lexical similarity, entailment (Mikolov et al., 2013a,b; Pennington et al., 2014; Bojanowski et al., 2016; Henderson and Popa, 2016).

We can, therefore, investigate whether the similarity spaces defined by word embeddings capture the notion of intervention similarity at work in long-distance dependencies. If they do, this means that they encode this core linguistic notion; if they don't this means that word embeddings semantic spaces capture association-based similarities based on world knowledge and textual co-occurrence, but not this more syntax-internal notion of intervention similarity.

3 The question

We investigate whether the popular notion of *word embeddings* and the notion of *vector space similarity* built on it are sensitive to the linguistic properties that are used to describe long-distance phenomena. These properties are the explanatory variables of the observed grammaticality judgments derived by intuitive or experimental acceptability judgments. If word embeddings encode the linguistic properties that explain grammaticality judgment in long-distance dependencies, then they should also be effective predictors of the grammaticality of these same sentences.

More precisely, let C and C' be the two elements linked by a long-distance dependency in sentence F. Let I be the intervener. Let $S(C, I)$ be a similarity score indicating how similar I is to C.[6] Let G_F be a score representing the grammaticality of F, as measured numerically by psycholinguistic controlled experiments. Intervention locality theory tells us that high $S(C, I)$ yields ungrammaticality. Then $S(C, I)$ is correlated to G_F.

We can encode this theory in vectorial space. Let w_C be the word embedding of C and w_I the word embedding of I. Let $s(w_C, w_I)$ be the similarity score S measured as a distance in vectorial space. Then $s(w_C, w_I)$ is correlated to G_F, if the similarity notion encoded in word embeddings is the similarity notion that has been shown to be active in long-distance dependencies. If instead word embeddings do not encode an intervention-sensitive notion of similarity, we should find no correlation.

For example, consider the weak island examples in Figure 2. Clearly, both the pair *(class, student)* and the pair *(professor, student)* are close in a semantic space that simply measures semantic field and association-based similarity. If however, word embeddings learn intervention-relevant notions of similarity, then *(professor, student*) should be more similar, since they are both animate, compared to *(class, student)*, a pair with a mismatch in animacy.

Note that it is crucial here to compute word embeddings in a way that does not encode grammatical, and especially syntactic, information in some other way, to control for effects of syntactic similarity. This could yield positive results for the wrong reasons. This is why we use syntax-lean vectors, as explained below, and not the more dynamic word embeddings calculated in the process of training a neural parser, for example, or a language model (Linzen et al., 2016; Bernardy and Lappin, 2017; Gulordava et al., 2018).

[5] Villata (2017, 8) summarizes that the relevant features have been identified as being morphosyntactic features that have the potential to trigger movement, such as [+Q], for *wh*-elements, [+R(el)], for the head of the relative clause, [+Top], for the elements in a topic position, [+Foc], for the focalized elements, and the [+N] feature associated with lexically restricted *wh*-elements (e.g., *which NP*).

[6] C and C' are fundamentally the same, so we will consider only C here.

Weak islands, ANIMACY MISMATCH

Quel cours te demandes-tu **quel étudiant** a apprécié?
[+Q,+N,-A] [+Q,+N,+A]

Which class do you wonder which student appreciated?

Weak islands, ANIMACY MATCH

Quel professeur te demandes-tu **quel étudiant** a apprécié?
[+Q,+N,+A] [+Q,+N,+A]

Which professor do you wonder which student appreciated?

Object relatives, NUMBER MATCH

Jules sourit à l' **étudiant** que l' **orateur** <étudiant>$_2$ endort <étudiant>$_1$ sérieusement depuis le début.

Jules smiles to the student who the speaker is putting seriously to sleep from the beginning.

Object relatives, NUMBER MISMATCH

Jules sourit aux **étudiants** que l' **orateur** <étudiants>$_2$ endorment <étudiants>$_1$ sérieusement depuis le début.

Jules smiles to the students who the speaker is putting seriously to sleep from the beginning.

Figure 2: The linguistic constructions and experimental materials

4 The experiments

In what follows, we describe the multiple steps necessary to construct the materials of our experiments. To verify our hypothesis, we need two sets of materials: the experimental measures reflecting the grammaticality of a sentence and the word embeddings to calculate a vector space of similarities. We describe these in turn. We refer to the sentences in Figure 2 as examples.

4.1 Materials

For grammaticality measures, we use the carefully controlled stimuli of three psycholinguistics experiments, kindly provided to us by S. Villata and J. Franck (Franck et al., 2015; Villata and Franck, 2016). The language studied is French. Subjects were not the same across the tasks. Stimuli are exemplified in Figure 2.

From Franck et al. (2015) we only consider the first experiment, comprising 24 experimental items crossing structure (object relative clauses vs. complement clauses) and the number of the object (singular vs. plural).[7]

The experimental data is constituted by on-line reading times (milliseconds). Interference is examined on the agreement of the verb in the subordinate clause. We use the reading time corresponding to the critical region, the verb following the intervener word, *endort* or *endorment* in our examples in Figure 2, as was done in the analysis of results in the original experiments. The results show a speed-up effect of number in number mismatches configurations.

From Villata and Franck (2016), we consider both experiments, both manipulating *wh*-islands. Experiment 1 manipulated the lexical restriction of the *wh*-elements (both bare vs. both lexically restricted), and the match in animacy between the extracted wh-element and the intervening *wh*-element (animacy match, where both are animate vs. animacy mismatch, where the extracted *wh*-element is inanimate and the intervening *wh*-element is animate). All verbs required animate subjects. Experiment 2 manipulated the lexical restriction of the wh-elements (both bare vs. both lexically restricted), and the reversibility of thematic roles (reversible vs. non-reversible). All *wh*-elements were animate.

The data collected are acceptability judgments collected off-line from several subjects, on a seven-point Likert scale.[8] The results show a clear effect of animacy match and reversibility of thematic role match for lexically restricted phrases and less so for bare *wh*-phrases.

Notice that these stimuli ensure that the effects, or, more importantly, null effects, that we might find are not limited to a single type of construction and lexical relation, since we test two very different sets of constructions. In the same spirit of testing for a wide set of effects, in one case, we look at effects expressed as offline acceptability, and in the other at online reading times.

4.2 Methods

Calculating the word and phrase vectors The pairs of words or phrases indicated in bold in the examples in Figure 2 were used to collect the vector-based similarity space.

For each of these words we recover a word embedding. We use French word embeddings, from

[7] All subject head nouns (e.g. *orateur*) were singular. Subjects and objects were all animate. An adverb followed by a locative phrase were added after the verb in order to measure potential spillover effects. All test sentences were grammatical with respect to subject-verb agreement. Each sentence was followed by a yes/no comprehension question that probed participants interpretation of the thematic relations in the sentence. Instructions encouraged both rapid reading and correctness in answering the questions (48 fillers, 72 subject).

[8] Subjects (42) were instructed that there were no time constraints. The stimulus set consisted of 32 experimental items that gave rise to 128 sentences and 132 fillers.

Facebook Research. 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).[9] Every word is represented as an n-grams of characters, for n training 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. This technique has been conceived to account for morphological similarities between words. Taking into consideration the fact that words may share morphological properties can improve the quality of the embeddings, and is important in a language like French, that has rich nominal and verbal inflectional morphology. The quality of a sample of these embedding vectors were checked by the two authors, proficient in French, by verifying that the words that are proposed as similar are consistent with intuition. Figure 3 shows the most similar words for two of the words whose word embeddings we calculated.

As shown in the examples in Figure 2, we need to measure the vector-based distance between phrases. Once the word vectors of individual words such as *quel* and *professeur*, are calculated, we calculate the embeddings of the noun phrases in which the single words combine, such as *quel professeur*. The vectorial representation of noun phrases is calculated by a composition operation. We used a simple vectorial sum. Since word embeddings are representations of lexical properties, we also report below results using only the bare head word of the noun phrase.

Calculating the similarity Once these vectors are calculated, we still have several options of which operator to use to calculate the distance between the vectors representing the two phrases C and I.

The similarity operators Beside the lexical specification of the vectors and their composition, the operator used to measure similarity also provides a dimension of experimental variation. The cosine is a well-known and efficient measure of vector similarity. It is based on a rescaling of the dot product of the vectors and it is a symmetric measure. It has been shown to capture associative and analogical semantic similarity in vector space (Mikolov et al., 2013a,b; Pennington et al., 2014; Bojanowski et al., 2016).

[9] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

POLICIER (policeman)	ETUDIANT (student)
cambrioleur (burglar)	enseignant (teacher)
kidnappeur (kidnapper)	professeur (professor)
chauffeur (driver)	chercheur (researcher)
criminel (offender)	doctorant (doctoral student)
détective (detective)	camarade (fellow)

Figure 3: Five most similar words for word *policier* and *étudiant*.

Once the distance between the vectors is calculated, in the final step, we correlate the calculated word embedding similarities with the psycholinguistic acceptability judgments.[10]

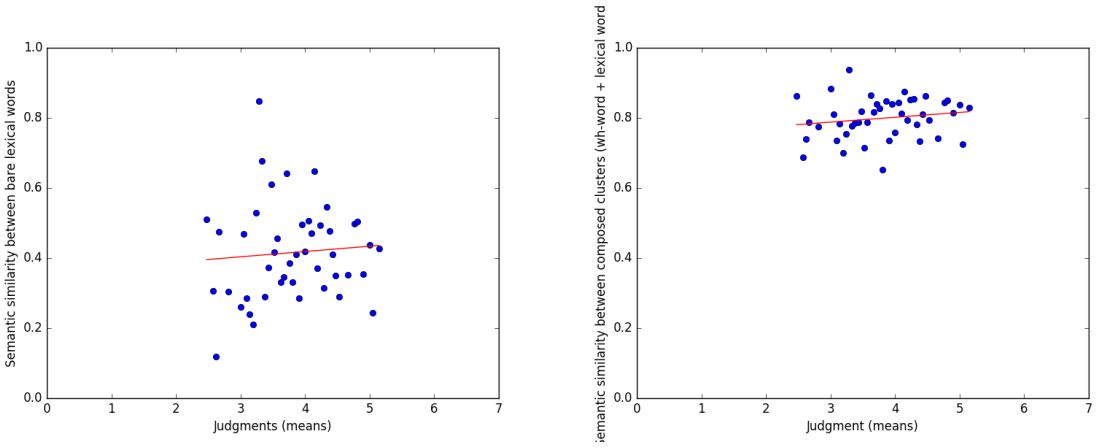
5 Results and discussion

Recall that in weak islands (see Figure 2), the expected outcome is an inverse proportionality between the two variables: the higher the semantic similarity, the stronger the interference, and consequently, the lower the average acceptability score of the sentence. In the case of object relative clauses (see Figure 2), we expected to observe a direct proportionality between the two variables: the higher the semantic similarity, the stronger the interference, and consequently the longer the average reading time devoted to the verb in the relative clause.

Results with the cosine operator Figures 4a and 4b show the (lack of) correlations between $s(w_C, w_I)$ and the grammaticality judgments of the experiments on weak islands, both with bare nouns and composed noun phrases. Figures 5a and 5b show the (lack of) correlations between $s(w_C, w_I)$ and the reaction times of the critical region, the verb, both with bare nouns and composed noun phrases, in object relative clauses. Regression values are shown in Table 1.

Results clearly show no correlations in all conditions. This is converging evidence that word embeddings do not represent the intervention notion of similarity, but they encode similarities based on associations and world knowledge. More explicitly, take the two examples of weak islands in Figure 2. Human judgments differentiate clearly the two sentences, the first being more acceptable than the second. In the first sentence, **Quel cours** te demandes-tu **quel étudiant** a apprécié? (*Which*

[10] The list of words and the detailed experimental results are given in the supplementary materials.



(a) Bare nouns (b) Composed phrases

Figure 4: Weak islands, cosine operator.

Op	Conf	Args	m	r	*p*
ss	WI	b1, b2	0.02	0.08	0.61
ss	WI	whp1, whp2	0.01	0.17	0.26
ss	OR	b1, b2	-5.83x10-5	-0.19	0.40
ss	OR	(b1,v),(b2,v)	2.42x10-5	0.08	0.72
as	WI	b1, b2	-0.97	-0.12	0.43
as	OR	b1, b2	-4x10-3	-0.22	0.33

Table 1: Regressions (m), correlations (Pearson r) and p-values. ss=semantic similarity (cosine); as=asymmetric similarity (lexical entailment); WI=weak island; OR=object relative clauses; b1/2=bare noun 1/2; whp1/2=*wh*-phrase 1/2; v=verb.

class do you wonder which student appreciated?), the two target words, in bold, do not match in animacy, hence the intervener does not block the long-distance relation as strongly as in the second sentence, **Quel professeur** te demandes-tu **quel étudiant** a apprécié? (*Which professor do you wonder which student appreciated?*), where they do. People are sensitive to this difference, even if *cours*, *professor* and *étudiant* are all words belonging to the same semantic field and closely connected by semantic association. The word embeddings we have tested here fail to capture this difference.

Analysis of results The lack of correlation prompts a more detailed analysis of the results. In particular, notice that in the experimental work a binary (not continuous) distinction – animate vs. inanimate, plural vs. singular – was manipulated and correlated to the acceptability and reaction times. We are, instead, requiring a correlation between similarity and acceptability in the animacy case and similarity and number in the reaction times. That is, we are imposing a stricter correspondence, which requires the level of similarity to continuously vary with all the experimental results. We verify then if weaker forms of correlation give us more positive results.

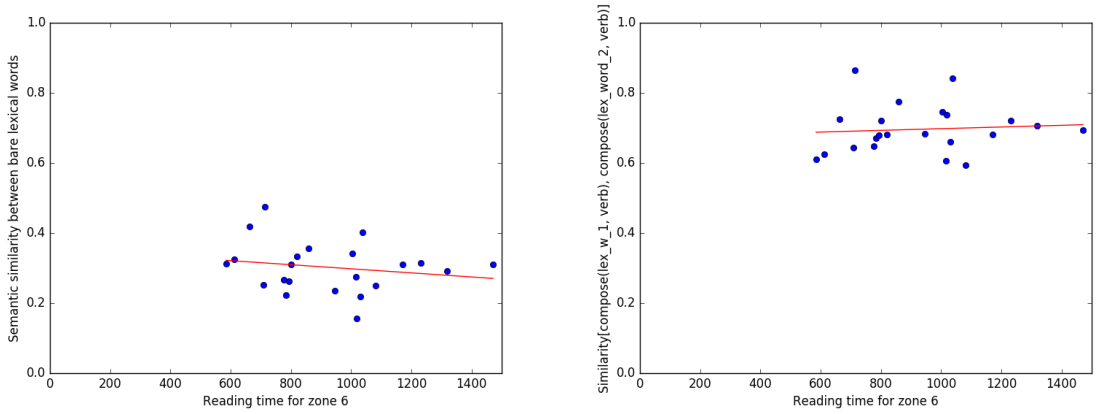
First of all, we can require the similarity measure to make only a binary distinction. For the experiment manipulating *animacy* in *wh*-islands, we do find the expected inverse correlation between mean similarity and mean acceptability depending on the value of the animacy factor. [11] For the experiment manipulating *number* in relative clauses, instead, we do not find the expected direct correlation between mean similarity and mean reading time depending on the value of the number factor.[12]

Another less stringent way of looking for correspondences is to take the manipulated binary factor into account, and verify if there is a partial correlation. In both cases, the correlation is weak.[13]

[11] Animate relative head (match condition): mean similarity=0.394, mean acceptability=3.65; inanimate relative head (mismatch condition): mean similarity=0.293, mean acceptability=4.00).

[12] Singular relative head (match condition): mean similarity=0.678, mean reading time=962.96; plural relative head (mismatch condition): mean similarity=0.705, mean reading time=896.03). Notice in fact, that the lack of correspondence could be even more basic, as the average similarity score for the number match condition is lower than for the number mismatch condition.

[13] A multiple regression of accuracy on animacy and similarity yields *accuracy= 0.46 anim=inanimate + 0.83 similarity + 3.33* with correlation coefficient 0.229; a multiple regression of reading times on number and similarity yields *reading times= 72.48 num=plural + 203.53 similar-*



(a) Bare nouns (b) Composed phrases

Figure 5: Object relative clauses, cosine operator.

Results with asymmetric operator It could also be pointed out that while the null results were confirmed across construction types (weak islands and object relatives), experimental methodologies (off-line grammaticality judgments and online reading times), only the cosine operator was used to calculate similarity. The two vectors that are being compared, w_C and w_I, correspond, linguistically to C and I above. It has been shown that, from a linguistic point of view, the grammaticality judgments differ depending on whether the feature set of C is properly included or properly includes I. If the features of C are a superset of the features of I, sentences are judged more acceptable (Rizzi, 2004). Independently of the exact details of the linguistic explanation, these fine-grained differences in grammaticality judgments suggest that it might be more appropriate to calculate similarity with an asymmetric operator.

The asymmetric measure we use here has been developed to capture the notion of entailment. It captures the idea that the values in a distributed semantic vector do not represent presence or absence of a property (true or false), but knowledge or lack of knowledge about a property of the referent entity of the noun whose meaning the vector represents: A entails B iff when I know A I know everything about B. This operator has been shown to learn the notion of hyponymy better than other methods (Henderson and Popa, 2016). [14]

Since this operator has so far only been applied to English, we need to develop the training and development sets for French. For our experiments, we translated all the word pairs from English to French.[15] We kept the same configurations of the training sets of word pairs, as described in the experiments by Henderson and Popa. The system uses these pairs coupled with the gold answer (1 if the entailment is true, 0 if it is not) to train on hyponymy-hypernymy relations. The data used for training are noun-noun word pairs that include positive hyponymy pairs, negative pairs consisting of different hyponymy pairs reversed, pairs in other semantic relations, and some random pairs.

We modify the operator (we use *unk dup*, $\circledcirc$), so that it does not to give us a binary decision (x entails y yes/no), but so that it outputs a real value, indicating how much x entails y, or rather how much x is asymmetrically similar to y.

With this operator, we produce the results shown in Figures 6a and 6b, for bare noun phrases.[16] Figure 6a shows the (lack of) correlations between $s(w_C, w_I)$ and the grammaticality judgments of the experiments on weak islands. Figure 6b shows the (lack of) correlations between

ity + 752.45, with correlation coefficient: -0.499.

[14] The operators are calculated by the following formula, where y, x are word embeddings vectors with length d, being projected in a different space.

$$\log(P(y \Rightarrow x)) \approx (\sigma(-(y-1)) \cdot \log \sigma(-(x-1)) + \sigma(-(-y-1)) \cdot \log \sigma(-(-x-1)))/d \quad (1)$$

The first dot product stands for the true-versus-unknown interpretation of the vectors and the second dot product represents the false-versus-unknown interpretation. σ is the logistic sigmoid function $\frac{1}{1+\exp(-x)}$, and the log and σ functions are applied componentwise.

[15] We use WordReference online multilingual dictionary, available at www.wordreference.com.

[16] Given the null results discussed below, we do not test another configuration, where we would have used the entailment operator on the composed noun phrase stimuli.

$s(w_C, w_I)$ and the reaction times of the critical region in object relative clauses. Regression values are shown in Table 1.

These results also confirm a lack of correlation. The convergence of these results is important as null effects are always hard to confirm and explain, and care must be taken to show that alternative explanations are not possible. In this case, all experiments, across constructions (weak island and object relative clauses), across type of noun phrase (bare or composed), across measurement method of the experimental dependent variable (off-line grammaticality judgments and online reaction times), and across operators (symmetric and asymmetric) show a consistent lack of correlation between measurements collected in experiments that manipulated the similarity of the elements, and the notion of similarity encoded in word embeddings.

This consistent lack of effect allows us to conclude that while current word embeddings, i.e. dictionaries in a multi-dimensional vectorial space, clearly encode a 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 in many human experiments to be at work and to be definitional in long-distance dependencies. They do not encode therefore this core notion of intervention similarity.

6 Related work

This work is situated in a rich body of computational research that attempts to establish the boundaries of what distributed semantic representations and neural networks can learn. These studies have concentrated on structural grammatical competence, exemplified by long-distance agreement, a task thought to require hierarchical, and not only linear, information. The first study, (Linzen et al., 2016), has tested recursive neural network (RNN) language models and found that RNNs can learn to predict English subject-verb agreement, if provided with explicit supervision. In a follow up paper, Bernardy and Lappin (2017) find that RNNs are better at long-distance agreement if they can use large vocabularies to form rich lexical representations to learn structural patterns. This finding suggests that RNNs learn syntactic patterns through rich lexical embeddings, based both on semantic and syntactic evidence. Gulordava et al. (2018) revisit previous work, and extend the work on long-distance agreement to four languages of different linguistic properties (Italian, English, Hebrew, Russian). They use the technique of developing counterfactual data, typical of theoretical and experimental work and already used for parsing in Gulordava and Merlo (2016) and train the system on nonsensical sentences. Their model makes accurate predictions and compares well with humans, thereby suggesting that the networks learn deeper grammatical competence.

On the linguistic and psycholinguistic side, this work contributes to the investigation of the formal encoding of long-distance dependencies, following the theoretical lines laid in the first formulation of intervention theory of long-distance dependencies (Rizzi, 1990), made gradual and more fine-grained in subsequent work (Rizzi, 2004), and verified experimentally in both sentence processing and acquisition (Franck et al., 2015; Villata and Franck, 2016; Friedmann et al., 2009).

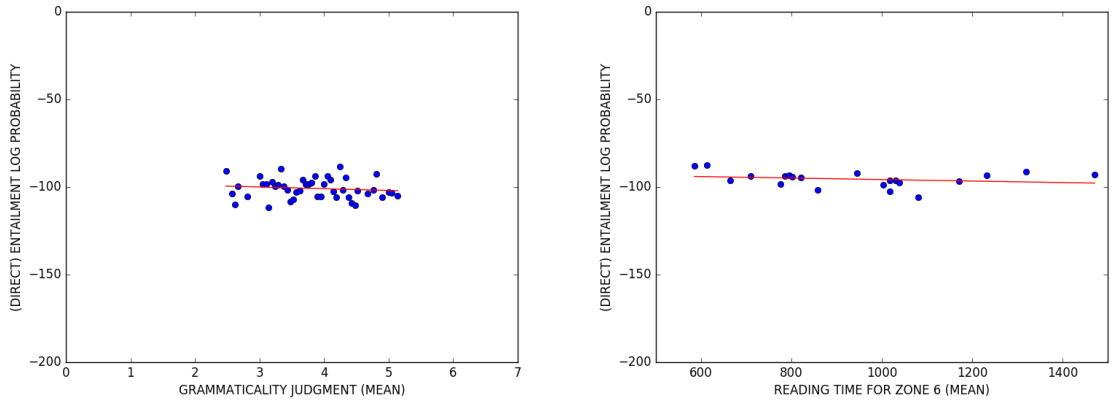
7 Conclusions

Human languages exhibit the ability to interpret discontinuous elements distant from each other in the string as if they were adjacent, but this long-distance relation can be disrupted by a similar intervening element. Speakers report lower acceptability and longer reading times. In this paper, we have presented results that show that the similarity spaces defined by one kind of word embeddings do not encode this notion of intervention similarity in long-distance dependencies.

Future work requires investigating more directly the grammatical aspects of the nature of the similar and dissimilar words in the embeddings and extend the experimentation to other kinds of vector spaces, a much larger dataset, and replication in more constructions and more languages.

8 Acknowledgments

We thank Julie Franck and Sandra Villata for sharing the data they have collected in their experiments, and James Henderson and Diana Nicoleta Popa for sharing with us their hyponymy detection script.



(a) Weak islands

(b) Object relatives

Figure 6: Asymmetric operator.

References

Flavia Adani. 2012. Some notes on the acquisition of relative clauses: new data and open questions. In *ENJOY LINGUISTICS! Papers offered to Luigi Rizzi on the occasion of his 60th birthday*, pages 6–13. CISCLPress.

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.

Jean-Philippe Bernardy and Shalom Lappin. 2017. Using deep neural networks to learn syntactic agreement. *Linguistic Issues in Language Technology*, 15(2):1–15.

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

Silke Brandt, Evan Kidd, Elena Lieven, and Michael Tomasello. 2009. The discourse bases of relativization: An investigation of young German and English-speaking children's comprehension of relative clauses. *Cognitive Linguistics*, 20(3):539–570.

Noam Chomsky. 2001. Derivation by phase. In Michael Kenstowicz, editor, *Ken Hale: A Life in Language*, pages 1–52. MIT Press, Cambridge,MA.

Ryan Cotterell and Hinrich Schütze. 2015. Morphological word-embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292, Denver, Colorado. Association for Computational Linguistics.

Katrin Erk. 2016. What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics*, 9(17):1–63.

John Rupert Firth. 1957. A synopsis of linguistic theory 1930-1955. In *Studies in linguistic analysis*, pages 1–32. Blackwell, Oxford.

Julie Franck, S. Colonna S., 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.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68:1–76.

Edward Gibson and Tessa Warren. 2004. Reading time evidence for intermediate linguistic structure in long-distance dependencies. *Syntax*, pages 55–78.

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.

Kristina Gulordava and Paola Merlo. 2016. Multilingual dependency parsing evaluation: a large-scale analysis of word order properties using artificial data. *Transactions of the Association for Computational Linguistics*.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

James Henderson and Diana Popa. 2016. A vector space for distributional semantics for entailment. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2052–2062, Berlin, Germany. Association for Computational Linguistics.

Aurélie Herbelot and Mohan Ganesalingam. 2013. Measuring semantic content in distributional vectors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 440–445, Sofia, Bulgaria. Association for Computational Linguistics.

Richard Kayne. 1989. Romance clitics,verb movement and PRO. *Linguistic Inquiry*, 22(4):647–686.

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.

Tal Linzen and Brian Leonard. 2018. Distinct patterns of syntactic agreement errors in recurrent networks and humans. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Paola Merlo. 2015. Evaluation of two-level dependency representations of argument structure in long-distance dependencies. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 221–230.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Joakim Nivre, Laura Rimell, Ryan McDonald, and Carlos Gómez Rodríguez. 2010. Evaluation of dependency parsers on unbounded dependencies. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 833–841, Beijing, China.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Laura Rimell, Stephen Clark, and Mark Steedman. 2009. Unbounded dependency recovery for parser evaluation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 813–821, Singapore. Association for Computational Linguistics.

Luigi Rizzi. 1990. *Relativized Minimality*. MIT Press, Cambridge, MA.

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.

Luigi Rizzi. 2013. Locality. *Lingua*, 130(1):69 – 86.

Marten van Schijndel and Tal Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Sandra Villata. 2017. *Intervention effects in sentence processing*. Ph.D. thesis, Universite de Geneve. Https://archive-ouverte.unige.ch/unige:101927.

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.

Ludwig Wittgenstein. (1953) [2001]. *Philosophical Investigations*. Blackwell Publishing.