Modeling Violations of Selectional Restrictions with Distributional Semantics

Emmanuele Chersoni, Adrià Torrens Urrutia, Philippe Blache, Alessandro Lenci

Modeling Violations of Selectional Restrictions with Distributional Semantics

Emmanuele Chersoni
Aix-Marseille University
emmanuelechersoni@gmail.com

Adrià Torrens Urrutia
Universitat Rovira i Virgili
adria.torrens@urv.cat

Philippe Blache
Aix-Marseille University
philippe.blache@univ-amu.fr

Alessandro Lenci
University of Pisa
alessandro.lenci@unipi.it

Abstract

Distributional Semantic Models have been successfully used for modeling selectional preferences in a variety of scenarios, since distributional similarity naturally provides an estimate of the degree to which an argument satisfies the requirement of a given predicate. However, we argue that the performance of such models on rare verb-argument combinations has received relatively little attention: it is not clear whether they are able to distinguish the combinations that are simply atypical, or implausible, from the *semantically anomalous* ones, and in particular, they have never been tested on the task of modeling their *differences in processing complexity*.
In this paper, we compare two different models of thematic fit by testing their ability of identifying *violations of selectional restrictions* in two datasets from the experimental studies.

1 Introduction

In recent years, Distributional Semantic Models (henceforth DSMs) have been at the core of one of the most active research areas in NLP, and have been applied to a wide variety of tasks. Among these, distributional modeling of selectional preferences (Erk et al., 2010; Baroni and Lenci, 2010) has been quite popular in computational psycholinguistics, since the similarity estimated by DSMs works very well for predicting the *thematic fit* between an argument and a verb. That is to say, the more the argument vector is similar to some kind of vector representation of the ideal filler of the verb slot (it can be either an abstract prototype, or a cluster of exemplars), the more the argument will satisfy the semantic requirements of the slot. The notion of thematic fit, as it has been proposed by the recent psycholinguistic research [1], is related to, but not totally equivalent to the classical notion of selectional preferences, since the former refers to a gradient compatibility between verb and role, whereas the latter conceives such compatibility as as boolean constraint evaluated on discrete semantic features (Lebani and Lenci, 2018).

The distributional models of thematic fit have been evaluated by comparing the plausibility scores produced by the models with human-elicited judgements (Erk et al., 2010; Baroni and Lenci, 2010; Greenberg et al., 2015; Santus et al., 2017), showing significant correlations. Moreover, they have been used to predict the composition and the update of argument expectations (Lenci, 2011; Chersoni et al., 2016), and for modeling reading times of experimental studies on complement coercion (Zarcone et al., 2013). However, an issue regarding their evaluation has not been addressed yet, i.e. their ability of capturing *different levels of implausibility*. [2]

Our processing system is sensitive to minimal variations in predictability between highly unpredictable word combinations, and such sensitivity has been shown to have an influence on reading times (Smith and Levy, 2013). Moreover, word combinations that are simply rare and/or unlikely and word combinations that are semantically deviant have been shown to have different consequences on processing complexity (Paczynski and Kuperberg, 2012; Warren et al., 2015).

[1] See McRae and Matsuki (2009) for an overview.

[2] A partial exception is the study on semantic deviance by Vecchi et al. (2011). However, they focus on the acceptability of adjectival phrases, rather than on selectional preferences.

From this point of view, thematic fit models represent an interesting alternative to the traditional probabilistic ones: they use distributional information about typical arguments to create an abstract representation of the "ideal" filler of the argument slot, and thus they are more capable of generalizing to the unseen. In other words, it does not matter if a specific verb-argument combination is attested in the training corpus of our system or not: its plausibility will still be computed on the basis of the similarity of the argument with the words that typically satisfy the requirements of the verb. It is important to stress that the inability to work with rare expressions has been for a long time a general point of criticism of statistical approaches to language, precisely because they could not explain why a given linguistic expression is not attested in the data (Vecchi et al., 2011).

In the present contribution, we take the first step toward the evaluation of thematic fit models on semantic anomaly detection. We set up a simple classification task on two datasets that have been recently introduced in the literature, and we test two different models on their ability to discriminate between a typical anomalous condition, i.e. **the violation of a selectional restriction**, and other highly unpredictable conditions.

2 Related Work

2.1 Distributional Semantic Models

All the DSMs rely on some version of the *Distributional Hypothesis* (Lenci, 2008), which can be stated as follows: *The semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B occur*.

The idea of analyzing meaning by measuring the similarity of distributional patterns turned out to be one of the most successful in the computational semantics research of the last two decades. Thanks to the improvements of automatic tools for language analysis and to the online availability of huge *corpora* of text, it has become easier and easier to automatically derive semantic representations of linguistic expressions in the form of *vectors* recording their contexts of occurrence. The closer two vectors in a distributional space, the more similar the meanings of the corresponding words.

Depending on the task, different definitions can be given to the notion of context: the contexts for a target word can be simply other words co-occurring within a sentence, within a word window with a fixed size or, as in our case, words that are syntactically related. In their most classical form, the so-called **Structured DSMs** use *syntactic relation: word* pairs as contexts to represent linguistic expressions. For example, *subject:baby, adverb: loudly* are possible contexts for the distributional representation of the verb *to cry*.

Since most DSMs of selectional preferences are structured and based on dependencies, also the models presented in this work will share the same features.

2.2 Thematic Fit and Distributional Semantics

Given a specific verb role-argument combination, the thematic fit task generally consists in predicting a value that expresses how well the argument fits the requirements of the role, e.g. how good is *burglar* as a patient for *arrest*. Since Erk et al. (2010), thematic fit models have been typically evaluated in terms of correlation of the model-derived scores with human-elicited judgements that have been collected for the purpose of psycholinguistic experiments (McRae et al., 1998; Ferretti et al., 2001; Padó, 2007; Hare et al., 2009). Erk and colleagues computed the fit of the candidate nouns by assessing their similarity with previously attested fillers of the respective roles. Going back to the previous example, if *burglar* is distributionally similar to the nouns of the entities that are typically *arrested*, then it should get a high score.

Baroni and Lenci (2010) similarly evaluated their Distributional Memory (DM) framework on the same task, adopting an approach that has became very popular in the literature: for each verb role, they built a single prototype vector by averaging the dependency-based vectors of its most typical fillers. The higher the similarity of a noun with a role prototype, the higher its plausibility as a filler for that role. Their model inspired several other studies: some of them tried to refine their DSM by using semantic roles-based vectors instead of dependency-based ones (Sayeed and Demberg, 2014; Sayeed et al., 2015)

or by using multiple prototypes, obtained through hierarchical clustering of the role fillers, in order to deal with verb polysemy (Greenberg et al., 2015).

An extension of the original model, introduced by Lenci (2011), has also been used to compute the dynamic update on the expectations for an argument filler, depending on how other roles have been filled in the previous part of the sentence (i.e., *engine* and *spelling* are both good patients for *to check*, but if the agent slot is filled by *mechanic*, then the former becomes a more predictable patient than the latter), and tested his system in a binary classification task on the subject-verb-object triples of the Bicknell dataset (Bicknell et al., 2010). More recently, Chersoni et al. (2016) integrated a similar mechanism of thematic fit computation in a more general model of semantic complexity, and obtained results comparable to Lenci (2011) on the same dataset.

Finally, Zarcone et al. (2013) made use of the notion of thematic fit in their study on complement coercion. Typically, we have a *complement coercion* when an event-selecting verb takes an entity-denoting NP as its direct object (i.e. *the author began the book*), so that a hidden verb has to inferred in order to satisfy the selectional restrictions of the verb (*the author began **writing** the book*). These authors computed the thematic fit for different verb-object combinations, corresponding to the experimental items used in the psycholinguistic experiments of McElree et al. (2001) and Traxler et al. (2002), and showed that the scores mirrored very closely the differences across conditions that were found in the above-mentioned studies. The coercion condition is particularly interesting for the present work, since it consists of an apparent violation of selectional restrictions. Therefore, the discrimination between actual violations and cases of complement coercion will be one of the tests for our models.

2.3 Experimental Evidence on Selectional Restrictions

Selectional restrictions can be defined as the set of semantic features that a verb requires of its arguments (Warren et al., 2015). Modular theories argued that they were represented in the lexicon, which was seen as a specialized module (Katz and Fodor, 1963; Fodor, 1983): it was generally assumed that the human comprehension system initially uses the knowledge available in such modules, and only later uses general world knowledge.

Since now there is evidence speaking against the modularity of the lexicon (Nieuwland and Van Berkum, 2006) and in favor of the access to world knowledge in the early stages of the comprehension process (McRae et al., 1998; McRae and Matsuki, 2009), it was questioned whether selectional restrictions have an independent reality, instead of being just part of a general world knowledge about events and participants (Hagoort et al., 2004; Kuperberg, 2007).

However, an EEG experiment by Pacyznski and Kuperberg (2012) showed that the processing difficulty of a sentence is affected differently by violation of selectional restrictions, with respect to simple event knowledge violation. The authors recorded ERPs on post-verbal Agent arguments as participants read passive English sentences, and they noticed that the N400 evoked by incoming animate Agent arguments violating event knowledge (e.g. *The bass was strummed by **the drummer***) was strongly attenuated when they were semantically related to the context (e.g. *the drummer* is related to a *concert*-type scenario). In contrast, semantic relatedness did not modulate the N400 evoked by inanimate Agent arguments that violated the preceding verbs animacy selection restrictions (e.g. *The bass was strummed by **the drum***). Such a result led the researchers to the conclusion that the two types of violations are actually distinct at the brain processing level.

Moreover, Warren et al. (2015) recently brought new evidence that the violation of a selectional restriction determines higher processing complexity than simple event implausibility. In an eye-tracking experiment, the authors compared the reading times between sentences in three different experimental conditions: a plausible condition (i.e. *The hamster explored a backpack*), an implausible condition with no violation of selectional restrictions (*The hamster lifted a backpack*) and an impossible condition with violation (*The hamster entertained a backpack*). Although the difference in human possibility ratings was not statistically significant between the last two conditions, eye-movements evidenced longer disruption in the violation condition compared to the other two. They concluded suggesting that selectional restrictions could actually be coarse-grained semantic features, derived by means of abstractions over

exemplar-type representations of events in memory. Violations of coarse-grained semantic features are likely to be detected earlier by the readers and cause more difficulty also in the later stages of processing, as they lead to such a degree of semantic anomaly that it becomes hard to build a coherent discourse model for the sentence (Warren and McConnell, 2007).

Most importantly, from a computational perspective, word combinations corresponding to the violations either of world knowledge (the implausible condition in Warren's data) or of selectional restrictions are not likely to be found in corpora of natural language data, and thus they cannot be distinguished on the basis of probabilistic methods. In our work we aim at testing the ability of thematic fit models to spot the difference and to assign different degrees of anomaly to the two conditions. The idea, intuitively, is that the degree of semantic anomaly goes hand in hand with an increase in processing complexity.

3 Experiments

For our experiments, we used two evaluation datasets: the sentences from the studies of Pylkkänen and McElree (2007) and Warren et al. (2015). The first study presented a magnetoencephalography experiment, with the goal of investigating the brain response to anomaly and to complement coercion, i.e. the case of a type clash between an event-selecting verb and an entity-denoting direct object. The experimental subjects were exposed to sentences in three different conditions: i) sentences with a typical verb-object combination (*The journalist wrote the article after his coffee break*); ii) sentences with a complement coercion (*The journalist began the article after his coffee break*); iii) sentences with a selectional restriction violation (*The journalist astonished the article after his coffee break*). This dataset is interesting for us because it will allow a direct comparison between violations of selectional restrictions and a similar phenomenon, the only difference being that a coercion involves the inference of a hidden verb (in the case of the example above, *writing*) that is not present in the linguistic input, leading to a sort of 'repair' of the violation. Discriminating between the two conditions is likely to be a difficult task.

The Warren dataset is the same of the study mentioned in Section 2.2. We are going to compare the items in the three conditions (plausible, implausible with no violation and impossible violation: see the examples in Section 2.2) of the experiment of Warren and colleagues, and we are particularly interested in the ability of the models to set the violation condition apart from the others. As declared by the authors themselves, they have built the sentences in a way than even the events described in the plausible condition are rare, or very unlikely. The test on this dataset will be particularly indicative of the performance of thematic fit models when they have to deal with different types of rare verb-argument combinations.

In both the datasets, we expect our thematic fit models to assign the lowest score to the violation condition, thus being able to distinguish between combinations that are simply unlikely and others that are really anomalous.

Datasets The Pylkkänen dataset is composed by 33 triplets of sentences, while the Warren dataset is composed by 30 triplets. We converted the experimental sentences in subject-verb-object triples. Here is one example from the Pylkkänen dataset (1) and one from the Warren dataset (2):

(1) a. *journalist-write-article* (typical)
b. *journalist-begin-article* (coercion)
c. *journalist-astonish-article* (violation)

(2) a. *hamster-explore-backpack* (plausible)
b. *hamster-lift-backpack* (implausible)
c. *hamster-entertain-backpack* (violation)

Before building our dependency-based DSM, we had to exclude three triplets from the Warren dataset since one or more words in the triplets had frequency below 100 in the training corpus. On the other hand, we have full coverage for the Pylkkänen dataset.

DSM We built a dependency-based DSM by using the data in the BNC corpus (Leech, 1992) and in the Wacky corpus (Baroni et al., 2009). Both the corpora were POS-tagged with the Tree Tagger (Schmid,

Verb and Role	Fillers
Agent of *to play*	actor, gamer, violinist
Agent of *to arrest*	cop, policeman, superhero
Patient of *to eat*	pizza, sandwich, ice-cream
Patient of *to shoot*	enemy, soldier, prey

Table 1: Verb roles and examples of fillers extracted by means of a corresponding syntactic relation.

1994) and parsed with the Maltparser (Nivre et al., 2006). [3] We extracted all the dependencies for the 20K most frequent words in the corpora, including the words of our datasets. Every co-occurrence between a target word and another context word in a given syntactic relation was weighted by means of Positive Local Mutual Information (Evert, 2004). [4] Given a target t, a relation r and a context word c occurring in the relation r with the target (e.g. $t = bark$, $r = sbj$, $c = dog$), we computed both their co-occurrence O_{trc}, and the expected co-occurrence E_{trc} under the assumption of statistical independence. The Positive Local Mutual Information (henceforth PLMI) is then computed as follows:

$$LMI(t, r, c) = log\left(\frac{O_{trc}}{E_{trc}}\right) * O_{trc} \tag{1}$$

$$PLMI(t, r, c) = max(LMI(t, r, c), 0) \tag{2}$$

Finally, each target word is represented by a vector of PLMI-weighted syntactic co-occurrences. Each contextual dimension corresponds to the co-occurrence of the target with a word in a given syntactic relation. For example, the vector of the verb *write-v* has dimensions such as *journalist-n:subj,article-n:obj* etc. [5]

Method As in Baroni and Lenci (2010), the thematic fit of a word for a given verb role is computed as the distributional similarity of that word with a *prototype representation* of the typical role filler. Such representation is obtained by averaging the vectors of the most typical fillers, i.e. words that are strongly associated with that verb-specific role. More concretely, the authors used syntactic functions to approximate thematic roles, and considered the most typical subjects of a verb as the fillers for the agent role, and the most typical objects as the fillers for the patient role. Typicality was measured by means of PLMI values: given a target verb t and a syntactic relation r, the typical fillers for the corresponding role were the 20 words with the highest PLMI association score with (t, r). Some examples of the extracted fillers are provided in Table 1. [6] Once built the prototype, the thematic fit of each candidate filler is assessed as the cosine similarity between the filler vector and the prototype itself.

For example, the prototype for the patient of *entertain-v* will be built out of the typical objects of the verb, such as *public, player* etc. Words that are distributionally similar to such fillers (i.e. *fan*) are likely to have a high thematic fit for the role.

Models In our experiments, we compared two different models of thematic fit. **B&L2010** is a 'classical' model of thematic fit, and it consists of a direct reimplementation of Baroni and Lenci (2010): since we are scoring sentences which differ for the degree of typicality of the verb-object combination, the scores assigned by this model will be the thematic fit scores θ of the object of each sentence given the verb and the patient role. In Equation 3, t is the target verb and c is a word occurring as an object (obj) of t:

$$\theta = \overrightarrow{c}|obj, \overrightarrow{t} \tag{3}$$

[3] We used the scripts of the DISSECT framework to build the distributional space (Dinu et al., 2013).

[4] As context words, we took into account only the 20K words of our target list, in order to limit the size of the distributional space.

[5] Obviously, including all the syntactic relations would have hugely increased the dimensionality of the vector space. Therefore, we took into account only the following relations: subject, direct and indirect object, prepositional complement. For each relation, we also considered its inverse: for example, the target *apple-v* has a dimension *eat-v:obj-1*, meaning that *apple* occurs as a direct object of *eat-v*.

[6] In the literature, 20 is a common choice for the number of fillers (Baroni and Lenci, 2010; Greenberg et al., 2015). Thus, we decided to keep this value for our experiments.

For example, the score of the sentence of the example 1a will be the thematic fit of the object *article-n* as a patient of *write-v*.

The second model is inspired by the proposal of Chersoni et al. (2016) who, instead of seeing the thematic fit as a simple measure of congruence between a predicate and an argument, considered it as a more general measure of the semantic coherence of an event. The global degree of semantic coherence is given by the product of the partial θ scores of all the event participants.

Similarly to Baroni and Lenci's model, each θ score is defined as the cosine similarity between an argument vector and the prototype vector for the slot, built as the centroid of its typical fillers. Once computed the partial θ scores, they are combined to find the global score θ_e.

$$\theta_e = \prod_{\overrightarrow{t},r,\overrightarrow{c} \in e} \theta(\overrightarrow{c}|r, \overrightarrow{t}) \tag{4}$$

where t is a target word in the event e [7], r is a syntactic relation and c is a context word occurring in the relation r with t (it is read as: the thematic fit score of c given the word t and the relation r).

For example, for the verb-argument triple of the example 1a, the three partial components of the final score would be: i) the thematic fit of the subject *journalist-n* as an agent of *write-v*; ii) the thematic fit of the object *article-n* as a patient of *write-v*; iii) the thematic fit of the object *article-n* as a co-argument of the subject *journalist-n*. [8]

The intuition of the authors was that the semantic coherence of an event does not depend simply on predicate-argument congruence scores, taken in isolation, but on a general degree of mutual typicality between all the participants. We will refer to this variant of the thematic fit model as **CBL2016**.

Task We evaluate the accuracy of the models in a classification task: for each triplet in the datasets, we compute the thematic fit scores for the subject-verb-object triples in the three conditions. We score a hit for a model each time it assigns the lowest score to the triple in the violation condition. The performance of both thematic fit models is compared to the one of a random baseline (since we have three different conditions, the accuracy is estimated to be 33.33%). We also use statistical tests to check in what measure the scores between the violation and the other conditions differ.

4 Results

The results of our experiments on the classification task are shown in Table 2 and Table 3. On the Warren dataset, the CBL2016 model performs extremely well, managing to assign the lowest thematic fit score to the violation condition in more than 80% of the triples of the dataset and reporting a highly significant advantage over the random baseline ($p < 0.001$) [9]. Although inferior in accuracy to the other model, B&L2010 manages as well to significantly outperform the baseline ($p < 0.05$). The Kruskal-Wallis test revealed a strong main effect of the condition on the scores assigned by both models (B&L2010: $\chi^2 = 20.502$, $p < 0.01$; CBL2016: $\chi^2 = 14.117$, $p < 0.01$). Post-hoc comparisons with the Wilcoxon rank sum test showed that, for both models, the scores differ significantly between the plausible and the violation condition and between the not plausible and the violation condition (in both cases, $p < 0.01$).

Model	Hits	Accuracy
Random	9/27	33.33%
B&L2010	18/27	66.66%
CBL2016	22/27	81.48%

Table 2: Accuracy scores for the Warren dataset.

[7] Keep in mind that, in the above-mentioned work, sentences are seen as linguistic descriptions of events and situations.

[8] The latter component was introduced because nouns, according to recent psycholinguistic studies (Hare et al., 2009; Bicknell et al., 2010), activate expectations about arguments typically co-occurring in the same events. In order to model the relationship between agents and patients of the same events, we introduced in our DSM the generic relation *verb* to link subjects and objects that tend to occur together, independently of the predicate.

[9] p-values computed with the χ^2 statistical test.

Model	Hits	Accuracy
Random	11/33	33.33%
B&L2010	21/33	63.63%
CBL2016	19/33	57.57%

Table 3: Accuracy scores for the Pylkkänen dataset.

These results are extremely relevant: although all the events of the Warren dataset have very low probabilities (for an explicit design choice of the authors), both the thematic models proved to be able to discriminate between events violating selectional restrictions and events that are simply unlikely (see also Figure 1, left side). They do not differ significantly for their ability to discriminate between the violation and the other conditions, as the violation consists of a mismatch of semantic features between the patient role of the verb and its filler (typically an animacy violation), and this information is available to both B&L2010 and CBL2016 in the form of an extremely low thematic fit for the patient. With respect to B&L2010, CBL2016 has also information on the thematic fit of the other event fillers. In theory, this should be an an advantage for distinguishing between the plausible and the not plausible condition: as it can be seen in Example 2, it is difficult to account for the difference in plausibility between a. and b. by only looking at the verb-patient combination. In practice, none of the models has assigned significantly different scores to the conditions a. and b., in line with the results of Warren et al. (2015), who also reported the absence of significant differences in reading times between plausible and not plausible sentences. This suggests that, for very rare events, different degrees of plausibility do not determine big changes in processing complexity, at least when selectional restrictions are not violated.

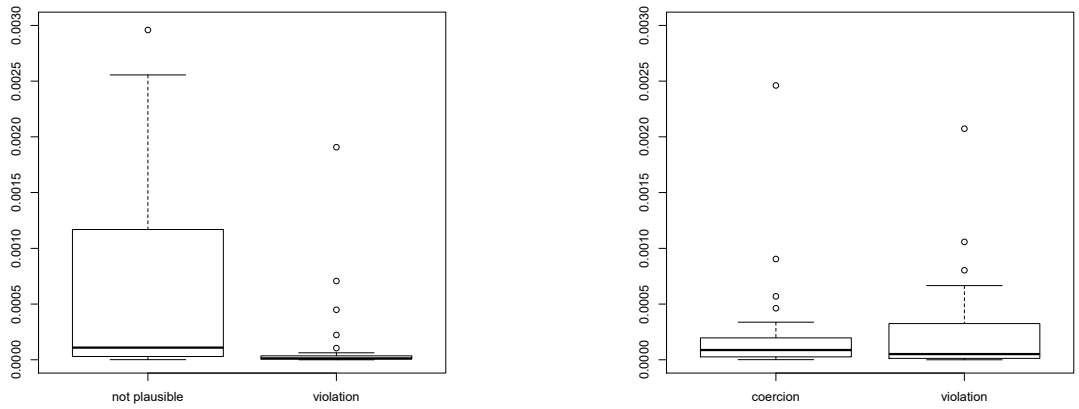

Figure 1: CBL 2016 score comparison between the NOT PLAUSIBLE and the VIOLATION condition on the Warren dataset (left) and between the COERCION and the VIOLATION condition on the Pylkkänen dataset (right).

As for the Pylkkänen dataset, both models were again able to outperform the random baseline on the classification task with a significant margin ($p < 0.05$) and, also on this dataset, the Kruskal-Wallis test showed a strong effect of the condition (B&L2010: $\chi^2 = 40.114$, $p < 0.001$; CBL2016: $\chi^2 = 13.804$, $p < 0.01$). The Wilcoxon test revealed that they are both efficient in discriminating between the typical and the other two conditions (B&L2010: $p < 0.001$ for both the typical-coercion and the typical-violation comparison; CBL2016: $p < 0.01$ for the same comparisons), but it revealed also an important difference: while B&L2010 assigns significantly higher scores to coerced sentences with respect to their counterparts containing violations ($p < 0.01$), CBL2016 fails to detect such a distinction ($p > 0.1$; see also Figure 1, right side). This result may seem surprising, since the less informed B&L2010 turns out to be the most efficient in detecting the fine-grained distinction between coercions and violations, simply on the basis of the typicality of the verb-patient argument combination.

A possible explanation is that the thematic fit was conceived in CBL2016 as a general index of se-

mantic coherence. If we limit ourselves to compute the fit between the event and the participants that are present in the linguistic input, it is not surprising that coercions and violations have similarly low coherence levels. After all, coercions can be described as violations of selectional restrictions that are repaired by inferring a hidden verb from the context (e.g. *writing* in *The journalist began the article*): since the model has no way to infer the hidden verb, it assigns a similarly low coherence score to the two experimental conditions.

5 Conclusion

In this paper, we have evaluated two thematic fit models in a classification task for the identification of violations of selectional restrictions. Our models had to deal with extremely rare word combinations (in the case of the Warren dataset) or to distinguish between violations and a similar phenomenon, i.e. complement coercion (in the case of the Pylkkänen dataset). On the Warren data, the performance of both models was very solid, clearly showing that they are able to discriminate between unlikely and anomalous inputs. Typically, such rare verb-argument combinations are not attested at all in corpora. We think this is a proof that the role characterization in thematic fit models allows generalizations on potential fillers that go well beyond the observable evidence. On the Pylkkänen dataset, the classical model by Baroni and Lenci (2010) manages to distinguish between coercion and violation, whereas the more recent model by Chersoni et al. (2016) does not. Still, the predictions of the latter could find some justification in the rationale behind its notion of thematic fit, and in the particular nature of the coercion phenomenon, describable as an apparent violation that is repaired by inferring a covert event.

More in general, the notion of thematic fit turns out to be very useful for modeling processing complexity, measured as in the experimental studies (mostly) in terms of processing times. Since thematic fit quantifies how a given argument fits a given semantic role, or a given event scenario, the low values correspond to situations in which it is extremely difficult to build a coherent semantic representation for the sentence. Given these promising results, future research should aim at building larger datasets to evaluate distributional models on anomaly detection tasks.

Another issue that deserves further investigation is the effect of the general discourse context on event plausibility, since contextual information in the current datasets is often limited to the other argument fillers. [10] As shown by studies like Warren et al. (2008), a context such as a fantasy world scenario can modulate the plausibility of an event and consequently the processing times, and the same could be true also for some specific real world scenarios (i.e. a psychiatric hospital, a circus etc.). Future efforts in modeling semantic anomalies have to take into account the acquisitions of the rich experimental literature on the topic, and to try to integrate as many as possible types of contextual manipulation in building new gold standards.

Acknowledgements

This work has been carried out thanks to the support of the A*MIDEX grant (nANR-11-IDEX-0001-02) funded by the French Government "Investissements d'Avenir" program.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A collection of Very Large Linguistically Processed Web-Crawled Corpora. In *Computational Linguistics*, 36(4): 673-721.

Marco Baroni and Alessandro Lenci. 2010. Distributional Memory: A General Framework for Corpus-based Semantics. In *Computational Linguistics*, 36(4): 673-721.

Klinton Bicknell, Jeffrey L Elman, Mary Hare, Ken McRae, and Marta Kutas. 2010. Effects of Event Knowledge in Processing Verbal Arguments. In *Journal of Memory and Language*, 63(4): 489-505.

Emmanuele Chersoni, Philippe Blache, and Alessandro Lenci. 2016. Towards a Distributional Model of Semantic Complexity. *Proceedings of the COLING Workshop on Computational Linguistics for Linguistic Complexity*.

[10] We thank one of the anonymous reviewers for pointing this out.

Georgiana Dinu, Nghia The Pham, and Marco Baroni. 2013. Dissect-Distributional Semantics Composition Toolkit. *Proceedings of ACL System Demonstrations*.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences. In *Computational Linguistics*, 36(4): 723-763.

Todd Ferretti, Ken McRae, and Andrea Hatherell. 2001. Integrating Verbs, Situation Schemas, and Thematic Role Concepts. In *Journal of Memory and Language*, 44(4): 516-547.

Stefan Evert. 2004. The Statistics of Word Cooccurrences: Word Pairs and Collocations. PhD Thesis.

Jerry Fodor. 1983. The Modularity of Mind. MIT Press.

Clayton Greenberg, Asad Sayeed, and Vera Demberg. 2015. Improving Unsupervised Vector-Space Thematic Fit Evaluation via Role-Filler Prototype Clustering. *Proceedings of NAACL-HLT*.

Peter Hagoort, Lea Hald, Marcel Bastiaansen, and Karl Magnus Petersson. 2004. Integration of Word Meaning and World Knowledge in Language Comprehension. In *Science*, 304(5669), 438–441.

Mary Hare, Michael Jones, Caroline Thomson, Sarah Kelly, and Ken McRae. 2009. Reading Time Evidence for Enriched Composition. In *Cognition*, 111(2), 151–167.

Jerrold J Katz and Jerry Fodor. 1963. The Structure of a Semantic Theory. In *Language*, 39(2), 170–210.

Gina R Kuperberg. 2007. Neural Mechanisms of Language Comprehension: Challenges to syntax. In *Brain Research*, 1146, 23–49.

Gianluca E Lebani and Alessandro Lenci. 2018. A Distributional Model of Verb-Specific Semantic Roles Inferences. In *Language, Cognition, and Computational Models*, edited by Thierry Poibeau and Aline Villavicencio. Cambridge University Press.

Geoffrey Neil Leech. 1992. 100 Million Words of English: The British National Corpus (BNC).

Alessandro Lenci. 2008. Distributional Semantics in Linguistic and Cognitive Research. *Italian Journal of Linguistics*, 20(1), 1–31.

Alessandro Lenci. 2011. Composing and Updating Verb Argument Expectations: A Distributional Semantic Model. *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics*.

Brian McElree, Matthew J Traxler, Martin J Pickering, Rachel E Seely, and Ray Jackendoff. 2001. Reading Time Evidence for Enriched Composition. In *Cognition*, 78(1), B17–B25.

Ken McRae, Micheal J Spivey-Knowlton, and Michael K Tanenhaus. 1998. Modeling the Influence of Thematic Fit (and Other Constraints) in Online Sentence Comprehension. In *Journal of Memory and Language*, 38(3), 283–312.

Ken McRae and Kazunaga Matsuki. 2009. People Use their Knowledge of Common Events to Understand Language, and Do So as Quickly as Possible. In *Language and Linguistics Compass*, 3(6), 1417–1429.

Mante S Nieuwland and Jos JA Van Berkum. 2006. When Peanuts Fall in Love: N400 Evidence for the Power of Discourse. In *Journal of Cognitive Neuroscience*, 18(7), 1098–1111.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Maltparser: A Data-Driven Parser-Generator for Dependency Parsing. *Proceedings of LREC*.

Martin Paczynski and Gina R Kuperberg. 2012. Multiple Influences of Semantic Memory on Sentence Processing: Distinct Effects of Semantic Relatedness on Violations of Real-World Event/State Knowledge and Animacy Selection Restrictions. In *Journal of Memory and Language*, 67(4), 426–448.

Ulrike Padó. 2007. The Integration of Syntax and Semantic Plausibility in a Wide-Coverage Model of Human Sentence Processing. PhD Thesis.

Liina Pylkkänen and Brian McElree. 2007. An MEG Study of Silent Meaning. In *Journal of Cognitive Neuroscience*, 19(11), 1905–1921.

Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. 2017. Measuring Thematic Fit with Distributional Feature Overlap. *Proceedings of EMNLP*.

Asad Sayeed and Vera Demberg. 2014. Combining Unsupervised Syntactic and Semantic Models of Thematic Fit. *Proceedings of CLIC.it*.

Asad Sayeed, Vera Demberg, and Pavel Shkadzko. 2014. An Exploration of Semantic Features in an Unsupervised Thematic Fit Evaluation Framework. In *Italian Journal of Computational Linguistics*.

Helmut Schmid. 1994. Part-of-Speech Tagging with Neural Networks. *Proceedings of COLING*.

Nathaniel J Smith and Roger Levy. 2013. The Effect of Word Predictability on Reading Time Is Logarithmic. In *Cognition*, 128(3), 302–319.

Matthew J Traxler, Martin J Pickering, and Brian McElree. 2002. Coercion in Sentence Processing: Evidence from Eye-Movements and Self-Paced Reading. In *Journal of Memory and Language*, 47(4), 530–547.

Eva Maria Vecchi, Marco Baroni, and Roberto Zamparelli. 2011. (Linear) Maps of the Impossible: Capturing Semantic Anomalies in Distributional Space. *Proceedings of the ACL Workshop on Distributional Semantics and Compositionality*.

Tessa Warren and Kerry McConnell. 2007. Investigating Effects of Selectional Restriction Violations and Plausibility Violation Severity on Eye-Movements in Reading. In *Psychonomic Bulletin and Review*, 14(4), 770–775.

Tessa Warren, Kerry McConnell and Keith Rayner. 2008. Effects of Context on Eye Movements when Reading about Possible and Impossible Events. In *Journal of Experimental Psychology*, 34(4).

Tessa Warren, Evelyn Milburn, Nikole D Patson, and Michael Walsh Dickey. 2015. Comprehending the Impossible: What Role Do Selectional Restriction Violations Play? In *Language, Cognition and Neuroscience*, 30(8), 932–939.

Alessandra Zarcone, Alessandro Lenci, Sebastian Padó, and Jason Utt. Fitting, not Clashing! A Distributional Semantic Model of Logical Metonymy. 2013. *Proceedings of IWCS*.