Is Structure Necessary for Modeling Argument Expectations in Distributional Semantics?

Emmanuele Chersoni
Aix-Marseille University
emmanuelechersoni@gmail.com

Enrico Santus
Singapore University of Technology and Design
esantus@mit.edu

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

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

Abstract

Despite the number of NLP studies dedicated to thematic fit estimation, little attention has been paid to the related task of composing and updating verb argument expectations. The few exceptions have mostly modeled this phenomenon with *structured* distributional models, implicitly assuming a similarly *structured* representation of events. Recent experimental evidence, however, suggests that human processing system could also exploit an unstructured "bag-of-arguments" type of event representation to predict upcoming input. In this paper, we re-implement a traditional structured model and adapt it to compare the different hypotheses concerning the degree of structure in our event knowledge, evaluating their relative performance in the task of the argument expectations update.

1 Introduction

An important trend of current research in linguistics and cognitive science aims at investigating the mechanisms behind anticipatory processing in natural language (Kamide et al., 2003; DeLong et al., 2005; Federmeier, 2007; Van Petten and Luka, 2012; Willems et al., 2015). It is, indeed, uncontroversial that our cognitive system tries to predict incoming input on the basis of prior information, and this strategy is probably crucial for dealing with the rapidity of linguistic interactions (Christiansen and Chater, 2016). By means of different experimental paradigms, several studies have focused on the role of event knowledge in the activation of expectations on verb arguments (McRae et al., 1998, 2005; Hare et al., 2009; Bicknell et al., 2010). During sentence processing, verbs (*arrest*) activate expectations on their typical argument nouns (*crook*), and nouns do the same for other arguments frequently co-occurring in the same events (*cop-crook*). The explanation proposed by these studies is that the human ability to anticipate the incoming input depends on general knowledge about events and their typical participants. This knowledge, stored in the semantic memory, 'reacts' to the linguistic stimulus: the more the processed information is coherent with a prototypical event scenario, the easier for the comprehension system is to constrain the range of the events potentially described by the sentence and to predict the upcoming sentence arguments.

According to one of the most influential accounts of event-based prediction, the event representation includes both the thematic roles and the lexical meanings of the arguments, as well as the relations between different roles. Therefore, there is the assumption of a *structural distinction* between the participants filling the roles (Kim et al., 2016) (i.e. some arguments are good *agents*, other are good *patients* etc.). Such an account has been challenged by the experimental evidence for a 'bag-of-arguments' mechanism of verb predictions, discussed by Chow et al. (2015). Their experiments with Event-Related Po-

tentials (ERPs) focused on the N400 component,[1] whose amplitude is generally interpreted as reflecting the predictability of a word in context (Kutas and Hillyard, 1984). One of their findings was that there is no significant difference in the N400 amplitude at the target verb in sentences like (1a) and (1b) (normal vs. *role-reversed* argument configuration):

(1) a. The restaurant owner forgot which **customer** the **waitress** had *served* during dinner yesterday.
b. The restaurant owner forgot which **waitress** the **customer** had *served* during dinner yesterday.

That is to say, even if different roles are assigned to *customer* and *waitress* in (1a) and (1b), this difference seems to have no impact on the N400 amplitude.

Given the lack of influence of the structural roles, in order to circumscribe their hypothesis (which we will henceforth refer to as the *bag-of-arguments hypothesis*), the authors set up another experiment, in which they tested whether the predictions could be influenced also by other co-occurring words (i.e., not necessarily arguments; we will refer to this possibility as the *bag-of-words hypothesis*). In order to carry out this test, they compared the amplitudes for sentences like (2a) and (2b) (argument substitution), finding a significantly smaller N400 component for the first sentence type.

(2) a. The exterminator inquired which **neighbor** the **landlord** had *evicted* from the apartment complex.
b. The neighbor inquired which **exterminator** the **landlord** had *evicted* from the apartment complex.

Chow et al. (2015) concluded that only the event *arguments* can influence predictions about a verb, and that arguments are represented in a sort of unstructured collection (i.e., *bag-of-arguments*). Therefore, according to them, predictions would be sensitive to the meaning of the arguments, but not to their structural roles, which are computed later. For example, the difference between typical agents and typical patients, according to this account, would not be included in the representation of an event.

In the last few years, a related issue has been debated in the field of distributional semantics, i.e. whether there is any added value in using structured representations of linguistic contexts over bag-of-words ones (e.g., contexts represented as co-occurrence windows). While structured models have been shown to outperform the latter in a number of semantic tasks (Padó and Lapata, 2007; Baroni and Lenci, 2010; Levy and Goldberg, 2014), some bag-of-word models proved to be extremely competitive, at least under certain parameter settings (Baroni et al., 2014). A recent paper by Lapesa and Evert (2017) explicitly addressed the question of whether using structured distributional semantic models is worth the effort, by comparing the performance of syntax-based and window-based distributional models on four different tasks. The authors showed that, even after extensive parameter tuning, the former have a significant advantage only in one task out of four (i.e., noun clustering). Interestingly, in the discussion they leave open the question of whether their results can generalize to linguistically challenging task such as the prediction of thematic fit ratings.

In this paper, we specifically investigate this point. The main questions we want to address are: what are the implications of the *bag-of-arguments hypothesis* for current models of thematic fit? More precisely, is it really necessary to have *structured* representations to carry out a thematic fit-related task, such as the argument expectation update?

In order to answer our questions, we implemented three models of argument expectations, adapting them to the above-mentioned hypotheses (i.e., *structured* and *unstructured*, the latter including both the *bag-of-arguments* and the *bag-of-words* hypothesis), and we compared their performance in a binary selection task.

[1]The N400, one of the most well-studied ERP components, is a negative-going deflection that peaks around 400 ms after the presentation of a stimulus word.

2 Related Work

One of the most influential distributional model of thematic fit was introduced by Baroni and Lenci (2010), who represented verb semantic roles with a *prototype vector* obtained by averaging the dependency-based vectors of the words typically filling those roles (i.e. the *typical fillers*). Within the Distributional Memory (DM) framework, which was based on syntactic dependencies, Baroni and Lenci used grammatical functions such as subject and object to approximate the thematic roles of agent and patient, and they measured role typicality by means of a Local Mutual Information score (Evert, 2004) computed between verb, arguments and syntactic relations. The basic assumption is that the higher the distributional similarity of a candidate argument with a role prototype, the higher its predictability as a filler for that role will be. As a gold standard, the authors used the human-elicited thematic fit ratings collected by McRae et al. (1998) and Padó (2007), and they evaluated the performance by measuring the correlation between these ratings and the scores generated by the model (as already proposed by Erk et al. (2010)).

Lenci (2011) later extended this 'structured-approach' to account for the dynamic update of the expectations on an argument, which depends on how other roles in the sentences are filled. For instance, given the agent *butcher* the expected patient of the verb *cut* is likely to be *meat*, while given the agent *coiffeur* the expected patient of the same verb is likely to be *hair*. By means of the same DM tensor, this study tested an additive and a multiplicative model (Mitchell and Lapata, 2010) to compose the distributional information coming from the agent and from the predicate of an agent-verb-patient triple (e.g., *butcher–cut–meat*), generating a prototype vector which represents the expectations on the patient filler, given the agent filler. The triples of the Bicknell dataset (Bicknell et al., 2010), which were used for the first time to evaluate such a model, are still today, at the best of our knowledge, the only existing gold standard for this type of task.

Although the 'structured-approach' to thematic fit was influential for a number of other works (Sayeed and Demberg, 2014; Sayeed et al., 2015; Greenberg et al., 2015; Sayeed et al., 2016; Santus et al., 2017), the task of modeling the update of the argument expectations has received relatively little attention. An exception is the work by Tilk et al. (2016), who trained a neural network on a role-labeled corpus in order to optimize the distributional representation for thematic fit estimation. Their model was also tested on the task of the composition and update of argument expectations, where it was able to achieve a performance comparable to Lenci (2011) on the triples of the Bicknell dataset.[2] Notice that both the models of Lenci (2011) and Tilk et al. (2016) necessarily rely on the hypothesis that the arguments are structurally distinct, since they are trained either on argument tuples containing fine-grained dependency information, or on sentences labeled with semantic roles.

Outside the specific area of study of thematic fit modeling, Ettinger et al. (2016) successfully used a type of unstructured representation for another sentence processing-related task, i.e. modeling N400 amplitudes with distributional spaces. The authors proposed a method based on word2vec (Mikolov et al., 2013) to build vector representations of sentence context, and to quantify the relation of an upcoming target word to the context. After training their word vectors on the Ukwac corpus (Baroni et al., 2009) with the Skip-Gram architecture, they modeled the mental state of a comprehender at a certain point of a sentence as the average of the vectors of the words in the sentence up to that point. The predictability of a target word in a sentence was measured as the cosine similarity between its vector, and the context-vector obtained by averaging the vectors of the preceding words. Ettinger and colleagues tested their method on the sentences used in the ERP study by Federmeier and Kutas (1999), in which three different conditions were defined, and they observed that the context-target similarity scores across conditions were following the same pattern of the N400 amplitudes of the original experiment. Thus, this work shows how data on N400 variations can be modeled even by means of vectors with minimal or no syntactic information.

[2]Chersoni et al. (2016) presented a research work testing a similar method on the Bicknell dataset. However, their model does not really update argument expectations on the basis of other arguments, computing instead a global score of semantic coherence for the entire event representation, on the basis of the mutual typicality between all the participants.

3 Experiments

Rationale. Baroni and Lenci (2010) computed the thematic fit for a candidate $filler$ (e.g., *policeman*) in an argument $slot$ (e.g., agent) of an $input$ lexical item (e.g., *arrest*) as the similarity score between the vector of the candidate $filler$ and a prototype of the typical $slot$ filler, built by summing the vectors for the top-*k* most typical fillers of $input$ for $slot$ (e.g., the typical agents for the *arrest*-event, such as *cop*, *officer*, *policewoman*, etc.). In this model, syntactic relations were used to approximate verb-specific semantic roles and to identify the most typical fillers. For example, the agent role is approximated by the subject relation, so that the typical $fillers$ for the agent $slot$ are the typical subjects of the $input$ verb. Similarly, the patient role is approximated by the object relation, so that the typical $fillers$ for the patient $slot$ are the typical objects of the $input$ verb.

We propose an extension of the model by Lenci (2011) and we interpret thematic fit as *the expectation of an argument* (i.e., what the prototype vector is meant to represent: $EX_{slot}(input)$), claiming that the update on expectations for a filler caused by new input (e.g. a verb combining with an agent) could be modeled by means of a function $f(x)$ that combines the prototypes built for every input:

$$EX_{slot}(\langle input_1, input_2\rangle) = f(EX_{slot_1}(input_1), EX_{slot_2}(input_2)) \quad (1)$$

where the function $f(x)$ is the sum or the pointwise multiplication between the prototype vectors. Once the expectations are calculated, the $filler$ fit for the $slot$ of $\langle input1, input2\rangle$ can be computed by measuring the similarity (e.g., by vector cosine) between the $filler$ and the expectations. As an example, if we want to estimate how likely is *burglar* as a patient of *the policeman arrested the...*, we build a prototype out of the vectors of typical objects co-occurring with the subject *policeman-n*, then we do the same for the vectors of typical objects of the verb *arrest-v*, and finally we combine the prototype vectors through $f(x)$, by either sum or multiplication. At this point, we can estimate the $filler$ fit by calculating the following similarity:

$$EX_{patient}(burglar|\langle police, arrest\rangle) = \\ sim(burglar, f(EX_{cooc\_patient}(policeman), EX_{patient}(arrest))) \quad (2)$$

Since distributional similarity is used as a measure of the predictability of a filler for a certain role, we expect that the thematic fit score of *burglar* for the patient $slot$ of $\langle policeman, arrested\rangle$ will be much higher than for *singer-n*. Indeed, *burglars* are more typical patients in this type of situation than *singers* are. Notice that while in Lenci (2011) the update function modified the association scores between the predicate and the fillers, in the present case $f(x)$ directly composes the prototype vectors associated with $\langle input1, input2\rangle$.

Models. In our experiments, we compared three different distributional semantic models (henceforth DSMs), all inspired by Lenci (2011): i) a structured model, which is similar to the one presented in Lenci (2011) (**DEPS**); ii) a variation of this system, modeling the *bag-of-arguments hypothesis* (**BOA**); iii) a baseline relying on the *bag-of-words hypothesis* (**BOW**). The key difference between our models is to be found in the selection of the fillers (see Table 1):

- **DEPS**: Similarly to Lenci's system, DEPS makes use of information on specific syntactic relations to select role fillers: the agent-role prototypes will be built out of the most typical subjects, the patient-role prototypes out of the most typical objects, and so on (see the last two rows of Table 1).[3] This means that not only the semantic information is taken in consideration (e.g. *policeman*), but also the thematic role of the filler (e.g. *sbj:policeman*, *obj:policeman*, etc.). Since dependencies

[3]Since the roles are approximated by syntactic relations identified by the parser (i.e. Malt-parser (Nivre and Hall, 2005)), their accuracy is subordinate to the accuracy of the parser. Ideally, we would expect clean lists of fillers for the typical subjects (agents) and objects (patients) of a verb, but – as it can be seen in Table 1 – this is not the case.

	Target	Fillers
BOW	steal-v	car-n, money-n, show-n, base-n, thief-n, good-n, item-n, property-n, someone-n, horse-n, limelight-n, vehicle-n, attempt-n, cattle-n, food-n, wallet-n, bike-n, identity-n, thunder-n, key-n
BOA	steal-v	show-n, money-n, car-n, base-n, food-n, thunder-n, march-n, limelight-n, horse-n, idea-n, key-n, wallet-n, heart-n, property-n, jewel-n, identity-n, cattle-n, body-n, purse-n, treasure-n
DEPS	steal-v (agent)	thief-n, someone-n, man-n, burglar-n, gang-n, money-n, robber-n, handbag-n, criminal-n, wallet-n, computer-n, thou-n, horse-n, equipment-n, boy-n, crook-n, disciple-n, cash-n, somebody-n, dog-n
DEPS	steal-v (patient)	show-n, money-n, car-n, base-n, food-n, thunder-n, march-n, limelight-n, horse-n, idea-n, key-n, wallet-n, heart-n, property-n, identity-n, cattle-n, jewel-n, body-n, purse-n, treasure-n

Table 1: Top-20 filler nouns for the word *steal-v* in our three models (for DEPS, we provide the fillers for the agent and the patient slot, while in the other models there is no distinction).

are used to filter fillers entering in the role representation, this model is the closest one to theories assuming *structured* event knowledge.

- **BOA**: Almost identical to the DEPS model, except for the fact that the most typical arguments are not bound to a specific syntactic slot. Indeed, according to Chow et al. (2015), the arguments of a verb like *to serve* (*customer*, *waitress*, *tray*, etc.) are represented like an *unstructured* collection. In this type of model, thus, the top-*k* typical fillers will include *all the strongly associated arguments*, abstracting away from the specific syntactic relation.

- **BOW**: In this baseline, the typical fillers are not arguments, but words typically co-occurring with the targets in a window of fixed width (possibly having no syntactic relation to the targets).

Going back to the previous example, the core idea of the DEPS model is that processing a sentence fragment like *the policeman arrested the...* leads to the activation only of the typical *patients* of such events, since the event knowledge is assumed to be structured. Therefore, the predictability of an argument is measured in terms of its similarity with the prototype built out of the activated patients.

On the other side, the BOA model assumes no distinction between the arguments (i.e., whether they are agents, patients, locations or others), and consequently the sentence fragment above would activate all the typical arguments of the verb *arrest*. This means that the predictability of an argument will be equivalent to its similarity with the prototype of a generic argument of the verb.

Finally, the BOW baseline has no notion of structure at all, not even the underspecified argument relation of the BOA model, and thus the prototypes of this model are just representations of the typical neighbors of the target words. It should be recalled at this point that a *bag-of-words* account of prediction was ruled out by the experimental results by Chow et al. (2015), since only arguments turned out to have an impact. Nonetheless, since we have chosen a *bag-of-words* model with a very narrow window (i.e., two words on the left and right of the target), BOW could also capture indirectly syntactic information (i.e., words frequently co-occurring with the targets within a narrow window are very likely to be also syntactically related to them). Therefore, we expect it to be a reasonably strong baseline.

Corpus and DSMs. Distributional information is derived from the concatenation of the British National Corpus (Leech, 1992) and of the Wacky (Baroni et al., 2009) corpus. Both were parsed with the Maltparser (Nivre and Hall, 2005). From this concatenation, we built a dependency-based DSMs, where the tuples are weighted by means of Positive Local Mutual Information (PLMI, Evert (2004)). Given the co-occurrence count O_{trf} of the target t, the syntactic relation r and the filler f, we computed the expected count E_{trf} (i.e., the simple joint probability of indipendent variables, corresponding to the product of the probabilities of the single events).[4]

[4]The DSM were built by means of the scripts of the DISSECT framework (Dinu et al., 2013)

The PLMI for each target-relation-filler tuple is computed as follows:

$$LMI(t, r, f) = log\left(\frac{O_{trf}}{E_{trf}}\right) * O_{trf} \tag{3}$$

$$PLMI(t, r, f) = max(LMI(t, r, f), 0) \tag{4}$$

Our DSM contains 28,817 targets (i.e., all nouns and verbs with frequency above 1000 in the training corpora), and all syntactic relations were included.[5] We also built a window-based DSM to extract co-occurrence information for the BOW model, counting only the co-occurrences between the nouns and the verbs of the list above within a word window of width 2.

Prototypes The prototypes of all models were built out of the vectors of the k most typical fillers for each model type, and we tested 10, 20, 30, 40, and 50 as values of k.[6]

As in previous studies, PLMI values were used as typicality scores: in the DEPS model, the typicality ranking of the fillers for a given role takes into account *only the fillers occurring in the corresponding syntactic slot* (e.g. the subject for the agent, the object for the patient etc.), whereas in the BOA model the typicality of a filler only depends on the PLMI score with the target, thus ignoring the type of syntactic relation.[7] As for the BOW baseline, the words used for building the prototype are simply co-occurring with the targets within a word window of width 2, and such co-occurrences have been PLMI-weighted as well.

Compositional Functions. The compositional functions that we used to combine the prototypes are the vector sum and the pointwise vector multiplication (Mitchell and Lapata, 2010). An important difference between the compositional functions lies in the fact that, while the sum retains the dimensions that are not shared by both prototype vectors, the multiplication sets them to zero those dimensions. This has an obvious impact on the computation of the cosine, as it could drastically reduce the number of dimensions on which the similarity score is computed.

Datasets and Evaluation. The models were tested on the datasets from the ERP experiments by Bicknell et al. (2010) and Chow et al. (2015).

The Bicknell dataset was introduced to test the hypothesis that the typicality of a verb direct object depends on the subject argument. With this purpose in mind, the authors selected 50 verbs, each paired with two agent nouns that significantly changed the scenario evoked by the subject-verb combination. They obtained typical patients for each agent-verb pair by means of production norms, and they used such data to generate triples where the patient was congruent with the agent and with the verb. For each congruent triple, an incongruent triple was generated as well, by combining each verb-congruent patient pair with the other agent noun, in order to have items describing atypical situations.

The final dataset is composed by 100 plausible-implausibile triples, which were used to build the sentences for a self-paced reading and for an ERP experiment. The subjects were presented with sentence pairs such as:

- *The journalist checked the spelling of the last report.* (plausible)
- *The mechanic checked the spelling of the last report.* (implausible).

Bicknell et al. (2010) reported shorter reading times and smaller N400 amplitudes for the plausible condition. The goal, for a thematic fit model of the argument expectations update, is to assign a higher

[5]We added the extra relation VERB, accounting for the link between typically co-occurring subjects and objects. An analogous relation was already in Baroni and Lenci (2010).

[6]The choice of the parameter range is in line with previous NLP studies on thematic fit (Sayeed et al., 2015; Greenberg et al., 2015).

[7]It goes without saying that using syntactic functions to identify the fillers of semantic roles is just an approximation. Nonetheless, the good performances reported by syntax-based thematic fit estimation systems suggest that, at least for agents and patients, such an approach is empirically justified.

cosine similarity score to the plausible triple, as in Lenci (2011). Moreover, Tilk et al. (2016) evaluated their systems on two different versions of this task, since the triple pairs can be created by combining either triples differing only for the agent, or triples differing only for the patient. Following the terminology from this latter study, we will refer to **Accuracy 1** meaning the accuracy of the models in scoring differing-by-patient triples, and to **Accuracy 2** meaning the accuracy in the classification of the differing-by-agent ones.

We also turned into similar triples the 50 verb-arguments combinations of the role reversal experiment by Chow et al. (2015), by creating triple pairs corresponding to the normal and to the role-reversed condition. For example, the sentences in Example (1) were turned into the form: *customer-n waitress-n serve-v* (normal) and *waitress-n customer-n serve-v* (role-reversed). Notice that we preserved the order in which the experimental subjects saw the arguments and the verb, with the latter at the end. Consequently, instead of composing the prototype vectors of the typical fillers of the patient role given an agent and a predicate, as we did for the Bicknell dataset, we derive the expectation vector for the verb from the composition of the prototypes of the typical predicates of the agent and of the patient.

The binary selection task is the same used with the Bicknell dataset, the only difference being that the goal for our models is to assign higher scores to the triples in the standard argument configuration (i.e., the expectation vector should be closer to the verb vector in the normal condition). Only the DEPS results are reported for the Chow dataset, because unstructured models assign exactly the same score to normal and role-reversed triples (independently of the order in which the prototypes of the head verb for each argument are created, the combined prototype will be the same). This is, of course, consistent with the report of the ERP experiment by Chow and colleagues, who found no differences in the N400 amplitudes elicited by the two sentence types.

The performance of the DEPS model on the Chow dataset is of particular interest, as the model has the structural information that is lacking in the other two. If DEPS has to reproduce the N400 pattern found by Chow and colleagues, the scores for the normal and for the role-reversed conditions should not differ significantly.

4 Results

In Table 2 and 3, we report the results for the three model types on the Bicknell dataset for the two kinds of prototype composition and $k = 20$. This latter value is the most common in the literature (Baroni and Lenci, 2010; Greenberg et al., 2015), and the one that gave us the highest accuracy scores.

The DEPS model is almost always the best performing one on the Bicknell dataset, with the exception of a single drop for the multiplicative model in the Accuracy 1 evaluation. The sum turned out to be the most efficient combination function in the majority of the models, and a possible explanation is that the application of multiplication to dependency-based prototype vectors led to sparsity problems. The results obtained by the DEPS Sum model are the highest ones, and the Accuracy 2 score for $k = 20$ is extremely close to the best performance reported in Lenci (2011) (73%).[8] The task of classifying differing-by-patient triples turns out to be harder, as the accuracies are lower and none of the models is significantly better than a random baseline (p-values were computed with the χ^2 statistical test)[9], whereas the Accuracy 2 scores of both the Deps Sum Model and the Deps Multiplication Model have a significant advantage (for both of them, $p < 0.05$).

We also carried out the Wilcoxon rank sum test on the scores generated by all models, and we found that the DEPS-sum model is the only one that manages to assign significantly different scores to the sentences in the two conditions ($W = 5660, p < 0.05$; for all the other models, $p > 0.1$) (see Figure 1). The BOA model was found instead to be worst performing one, even lower than the BOW baseline, and often the recorded accuracy scores are very close to a random baseline. Also, the differences between conditions were far from significance in any of the versions of the model.

[8] The only result available for comparison in the literature is the one obtained by Lenci (2011) on the Accuracy 2 task, since the evaluation of Tilk et al. (2016) was carried out on a way smaller subset of the Bicknell dataset (64 triples).

[9] Also the Accuracy 1 scores reported by Tilk et al. (2016) confirm the higher difficulty of this version of the task.

Model	Sum	Multiplication
BOW	59%	57%
BOA	53%	59%
DEPS	**62%**	56%

Table 2: Accuracy 1 on Bicknell (100% coverage) for $k = 20$

Model	Sum	Multiplication
BOW	60%	56%
BOA	58%	57%
DEPS	**72%**	68%

Table 3: Accuracy 2 on Bicknell (100% coverage) for $k = 20$

Figure 2 shows the performance variation of the Sum models on Bicknell dataset, while varying the number of fillers used to build the prototype. At a glance, we can observe that DEPS models achieve higher accuracies with fewer fillers. This is kind of expected, since the good performances of such models are likely to be due to a more restrictive selection of the fillers. With higher values of k, the selection of more weakly-related fillers is probably introducing noise in the prototype. On the other hand, BOA models slightly improve when more fillers are used, but in general their performance is almost always equivalent to BOW models. This indicates, in our view, that the underspecified dependency relation of the BOA model is insufficient to build a precise representation of the expectations on an upcoming argument, unless a larger number of fillers is taken into account. Moreover, even if typical arguments are selected by virtue of a dependency relation, the absence of information on the dependency type makes these models essentially equivalent to window-based ones. Finally, as for the difference between the scores in the two conditions, the Wilcoxon rank sum test returns a significant difference only for the DEPS Sum model with $k = 10, 20$ (in both cases, $p < 0.05$).

Concerning the performance of DEPS on the Chow dataset, it can be seen in Table 4 that the system identifies the triple in the normal condition with a level of accuracy between 62% and 68%. The scores are quite steady, independently from k, and again, the Sum models are generally performing better (but never significantly better than a random baseline: for all parameter settings, $p > 0.05$).

Interestingly, after applying the Wilcoxon rank sum test, it turns out that the differences between the assigned scores never differ significantly between the normal and the argument reversal condition (for all values of k, $p > 0.05$; see also Figure 1). This result is coherent with the outcome of the role-reversal experiment by Chow and colleagues, who found no difference between the N400 elicited by the two sentence types. In other words, structural information does not help in predicting the upcoming verb.

k	Sum	Multiplication
10	68%	66%
20	62%	64%
30	64%	62%
40	64%	62%
50	66%	62%

Table 4: Accuracy on Chow (100% coverage) for the DEPS model for different values of k

In the same way as the unstructured representations used by Ettinger et al. (2016), our models show that the distributional similarity between a target and its context (a structured one, in our case) can accurately reflect the N400 amplitude patterns found in the experimental studies. Notice however that only a model based on the notion of a structured event knowledge was able to mirror the patterns of both the studies of Bicknell et al. (2010). Together with the better performance reported on the Bicknell dataset for the binary classification task, these results suggest that the presence of structural information is an advantage for distributional models of thematic fit.

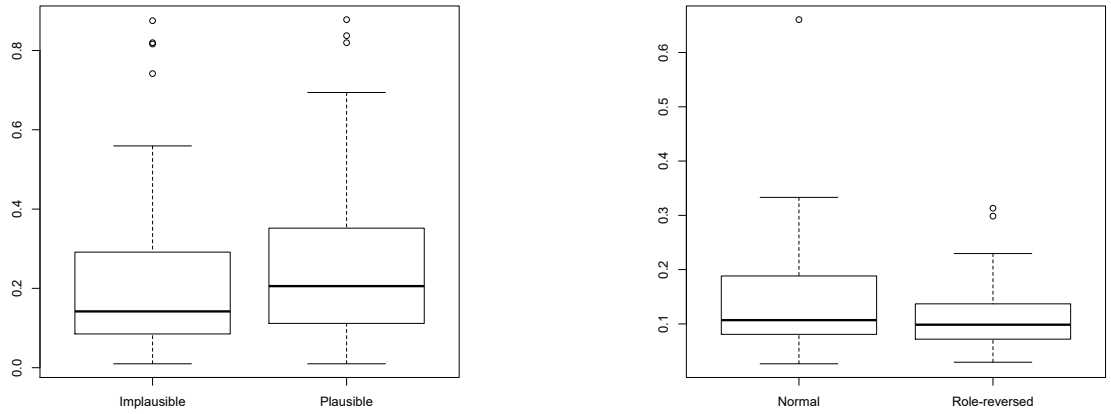

Figure 1: Cosine scores assigned by DEPS-sum model ($k = 20$) for the Bicknell dataset (left) and for the Chow dataset (right).

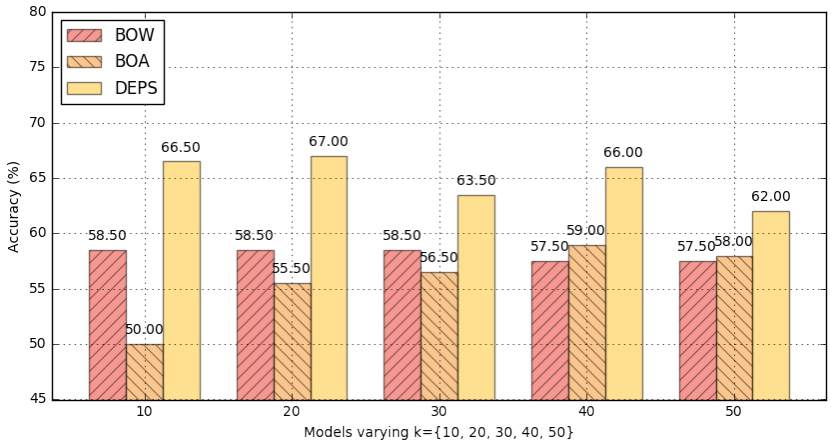

Figure 2: Average of Accuracy 1 and Accuracy 2 on Bicknell dataset, for all the Sum models for different values of k

5 Conclusions

In this paper, we addressed the question of whether structured information is necessary to model the argument expectation update. With this purpose in mind, we have implemented a traditional system for composing and updating thematic fit estimations (Lenci, 2011) and we adapted it to model both structured and unstructured representations (the latter including both the *bag-of-arguments* and the *bag-of-words* hypotheses). We compared the performance of these models on the binary selection task of the argument expectation update and on their ability to replicate the experimental results from the studies by Bicknell et al. (2010) and Chow et al. (2015).

Our results show that structured models perform better in a task of composing and updating argument expectations, and can reproduce the ERP results reported for both datasets. On the other hand, the *bag-of-arguments* model had lower scores in the classification task on the Bicknell dataset, and it was not able to discriminate between the plausible and the implausible condition.

It should be recalled that the *bag-of-arguments* hypothesis was proposed to account for the results of an experiment on initial verb predictions, where the participants could see the verb only at the end (see Examples 1 and 2). In absence of any cue facilitating the mapping between arguments and syntactic positions (consider also that the arguments in the dataset do not differ by animacy), it is reasonable to

hypothesize a delay in the assignment of the thematic roles. Moreover, as already pointed out by Kim et al. (2016), the N400 component is generally not sensitive to the implausibility derived by thematic role reassignments, but the presence of event knowledge violation in such cases can be signaled by other ERP components.[10] In sum, the idea of a structure in the event knowledge does not seem to be incompatible with the findings of Chow and colleagues, since our structured DSMs replicated the lack of significant differences between normal and role-reversed sentences. On the other side, models with no structural information struggle in modeling the results of datasets where the items differ for their context-sensitive argument typicality, like the one from Bicknell et al. (2010).[11]

The performance of the DEPS model also complies with the conclusions of Ettinger et al. (2016), which showed how DSMs could be used to reproduce the N400 variations. Such a component is known to be tied to the general semantic relatedness of a target word to its sentential context, and not to syntactic anomalies.[12] From this point of view, it is interesting that our structured models, despite their coherence with the ERP results by Chow et al. (2015), are still able to distinguish the sentence in the normal condition from the role-reversed one with an accuracy always above 60%. Future research could explore in which measure thematic fit models can be sensitive to differences between syntactically-composed representation. Finally, with reference to Lapesa and Evert (2017), our results make the expectation update task a good candidate for being among those that clearly benefit from using fine-grained syntactic information, as it seems to require knowledge about the relation types and about the interdependencies between participants.

Future works might aim at comparing these model types on other NLP tasks, to check how many of them effectively take advantage from structured representations. For the moment, we can conclude that structure is an important added value for thematic fit models.

Acknowledgments

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.
We would like to thank the anonymous reviewers for their comments and for their helpful suggestions.

References

Baroni, M., S. Bernardini, A. Ferraresi, and E. Zanchetta (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-crawled Corpora. *Language Resources and Evaluation 43*(3), 209–226.

Baroni, M., G. Dinu, and G. Kruszewski (2014). Dont Count, Predict! A Systematic Comparison of Context-counting vs. Context-predicting Semantic Vectors. In *Proceedings of ACL*.

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

Bicknell, K., J. L. Elman, M. Hare, K. McRae, and M. Kutas (2010). Effects of Event Knowledge in Processing Verbal Arguments. *Journal of Memory and Language 63*(4), 489–505.

[10] Kim et al. (2016) point out that role-reversed sentences typically elicit en enhanced P600 components compared to their plausible counterparts (see also the experiments in Kim and Osterhout (2005) and Kim and Sikos (2011)).

[11] The 'bag-of-arguments' mechanism described by Chow et al. (2015) actually concerns a very early stage of the comprehension process. Moreover, in a later response article, Chow et al. (2016) brought evidence that verb predictions become sensitive to structural roles of the arguments if more time is available for prediction.
We thank one of our reviewers for pointing this out.

[12] The study of Chow et al. (2015) is an example of experimental evidence for this claim, but see also the results from Fischler et al. (1985) on the N400 insensitivity to the introduction of negations.

Chersoni, E., P. Blache, and A. Lenci (2016). Towards a Distributional Model of Semantic Complexity. *Proceedings of the COLING Workshop on Computational Linguistics for Linguistic Complexity*.

Chow, W.-Y., S. Momma, C. Smith, E. Lau, and C. Phillips (2016). Prediction as memory retrieval: timing and mechanisms. *Language, Cognition and Neuroscience 31*(5), 617–627.

Chow, W.-Y., C. Smith, E. Lau, and C. Phillips (2015). A 'Bag-of-arguments' Mechanism for Initial Verb Predictions. *Language, Cognition and Neuroscience (Advance online publication) 31*(5), 577–596.

Christiansen, M. H. and N. Chater (2016). The Now-or-Never Bottleneck: A Fundamental Constraint on Language. *Behavioral and Brain Sciences 39*.

DeLong, K. A., T. P. Urbach, and M. Kutas (2005). Probabilistic Word Pre-activation During Language Comprehension Inferred from Electrical Brain Activity. *Nature Neuroscience 8*(8), 1117–1721.

Dinu, G., N. T. Pham, and M. Baroni (2013). DISSECT-DIStributional SEmantics Composition Toolkit. In *Proceedings of the ACL System Demonstrations*.

Erk, K., S. Padó, and U. Padó (2010). A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics 36*(4), 723–763.

Ettinger, A., N. H. Feldman, P. Resnik, and C. Phillips (2016). Modeling N400 Amplitude Using Vector Space Models of Word Representation. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, pp. 1445–1450.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph. D. thesis.

Federmeier, K. D. (2007). Thinking Ahead: The Role and Roots of Prediction in Language Comprehension. *Psychophysiology 44*(4), 491–505.

Federmeier, K. D. and M. Kutas (1999). A Rose by any Other Name: Long-term Memory Structure and Sentence Processing. *Journal of Memory and Language 41*(4), 469–495.

Fischler, I., D. G. Childers, T. Achariyapaopan, and N. W. Perry (1985). Brain Potentials During Sentence Verification: Automatic Aspects of Comprehension. *Biological Psychology 21*(2), 83–105.

Greenberg, C., A. B. Sayeed, and V. Demberg (2015). Improving Unsupervised Vector-space Thematic Fit Evaluation via Role-filler Prototype Clustering. In *Proceedings of HLT-NAACL*.

Hare, M., M. Jones, C. Thomson, S. Kelly, and K. McRae (2009). Activating Event Knowledge. *Cognition 111 2*, 151–67.

Kamide, Y., G. T. Altmann, and S. L. Haywood (2003). The Time-course of Prediction in Incremental Sentence Processing: Evidence from Anticipatory Eye Movements . *Journal of Memory and Language 49*(1), 133 – 156.

Kim, A. E., L. D. Oines, and L. Sikos (2016). Prediction During Sentence Comprehension is More than a Sum of Lexical Associations: the Role of Event Knowledge. *Language, Cognition and Neuroscience 31*(5), 597–601.

Kim, A. E. and L. Osterhout (2005). The Independence of Combinatory Semantic Processing: Evidence from Event-related Potentials. *Journal of Memory and Language 52*(2), 205–225.

Kim, A. E. and L. Sikos (2011). Conflict and Surrender During Sentence Processing: An ERP Study of Syntax-semantics Interaction. *Brain and Language 118*(1), 15–22.

Kutas, M. and S. A. Hillyard (1984). Brain Potentials During Reading Reflect Word Expectancy and Semantic Association. *Nature 307*, 161–163.

Lapesa, G. and S. Evert (2017). Large-scale Evaluation of Dependency-based DSMs: Are They Worth the Effort? In *Proceedings of EACL*.

Leech, G. (1992). 100 Million Words of English: the British National Corpus (BNC). *Language Research 28*(1), 1–13.

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

Levy, O. and Y. Goldberg (2014). Dependency-Based Word Embeddings. *In Proceedings of ACL*.

McRae, K., M. Hare, J. L. Elman, and T. Ferretti (2005). A Basis for Generating Expectancies for Verbs from Nouns. *Memory & Cognition 33*(7), 1174–1184.

McRae, K., M. J. Spivey-Knowlton, and M. K. Tanenhaus (1998). Modeling the Influence of Thematic Fit (and Other Constraints) in On-line Sentence Comprehension. *Journal of Memory and Language 38*, 283–312.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Mitchell, J. and M. Lapata (2010). Composition in Distributional Models of Semantics. *Cognitive Science 34*(8), 1388–1429.

Nivre, J. and J. Hall (2005). Maltparser: A Language-independent System for Data-driven Dependency Parsing. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, pp. 13–95.

Padó, S. and M. Lapata (2007). Dependency-Based Construction of Semantic Space Models. *Computational Linguistics 33*(2), 161–199.

Padó, U. (2007). *The Integration of Syntax and Semantic Plausibility in a Wide-coverage Model of Human Sentence Processing*. Ph. D. thesis.

Santus, E., E. Chersoni, A. Lenci, and P. Blache (2017). Measuring thematic fit with distributional feature overlap. In *Proceedings of EMNLP*.

Sayeed, A. and V. Demberg (2014). Combining Unsupervised Syntactic and Semantic Models of Thematic Fit. In *Proceedings of CLIC-IT*.

Sayeed, A., V. Demberg, and P. Shkadzko (2015). An Exploration of Semantic Features in an Unsupervised Thematic Fit Evaluation Framework. *Italian Journal of Linguistics*.

Sayeed, A., C. Greenberg, and V. Demberg (2016). Thematic Fit Evaluation: an Aspect of Selectional Preferences. In *Proceedings of the ACL Workshop on Evaluating Vector-Space Representations for NLP*.

Tilk, O., V. Demberg, A. B. Sayeed, D. Klakow, and S. Thater (2016). Event Participant Modelling with Neural Networks. In *Proceedings of EMNLP*.

Van Petten, C. and B. J. Luka (2012). Prediction During Language Comprehension: Benefits, Costs, and ERP Components. *International Journal of Psychophysiology 83*(2), 176–190.

Willems, R. M., S. L. Frank, A. D. Nijhof, P. Hagoort, and A. Van den Bosch (2015). Prediction During Natural Language Comprehension. *Cerebral Cortex*, bhv075.