Towards a Distributional Model of Semantic Complexity

Emmanuele Chersoni, Philippe Blache, Alessandro Lenci

Towards a Distributional Model of Semantic Complexity

Emmanuele Chersoni
Aix-Marseille University
emmanuelechersoni@gmail.com

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

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

Abstract

In this paper, we introduce for the first time a Distributional Model for computing semantic complexity, inspired by the general principles of the Memory, Unification and Control framework (Hagoort, 2013; Hagoort, 2016). We argue that sentence comprehension is an incremental process driven by the goal of constructing a coherent representation of the event represented by the sentence. The composition cost of a sentence depends on the *semantic coherence* of the event being constructed and on the *activation degree* of the linguistic constructions. We also report the results of a first evaluation of the model on the Bicknell dataset (Bicknell et al., 2010).

1 Introduction

The differences in semantic processing between *typical* and *atypical* sentences have recently attracted a lot of attention in experimental linguistics. Consider the following examples:

(1) a. *The musician plays the flute in the theater.*
b. *The gardener plays the castanets in the cave.*
c. ** The nominative plays the global map in the pot.*

Since the early work of Chomsky (1957) and the introduction of the notion of acceptability, linguistic theory has mostly focused on the contrast between (1c) and the former two. The last sentence violates the combinatorial constraints of the lexical items, and that is the reason why, although (1c) is syntactically well-formed, we are not able to build any coherent representation for the situation it expresses. Investigations on event-related potentials (ERP)[1] brought extensive evidence that sentences like (1a) and (1b), despite being both semantically acceptable, have a different cognitive status: sentences such as (1b), including possible but unexpected combinations of lexemes, evoke stronger N400 components[2] in the ERP waveform than sentences with non-novel combinations, like (1a)(Baggio and Hagoort, 2011).

Although there are different interpretations of the N400 effect,[3] there is general agreement among researchers that it is a brain signature of *semantic complexity*, that can be reinforced at the syntactic level (the *syntactic boost* effect; see (Hagoort, 2003)): novel and unexpected combinations are more complex and require larger cognitive efforts for processing. An open question is what are the factors determining the semantic complexity of sentence comprehension. Baggio et al. (2012) claim that the real issue about compositionality and open-ended productivity is the *balance between storage and computation*. Productivity entails that not everything can be stored. However, the N400 effect suggests that there is a large amount of stored knowledge in semantic memory about event contingencies and concept combinations.

[1] An event-related potential is an electrophysiological response of the brain to a stimulus.

[2] The N400 is a negative-going deflection that peaks around 400 milliseconds after presentation of the stimulus.

[3] See, for example, Kutas and Federmaier (2000) and Van Berkum et al. (2005) for the 'feature pre-activation hypothesis'; and Baggio et al. (2012) for an interpretation of the N400 amplitude as a consequence of the cost of semantic unification. In a connectionist framework, McClelland (1994) and, more recently, Rabovsky and McRae (2014) have advanced the hypothesis that N400 amplitudes correlate with implicit prediction error in the semantic memory.

Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity,
pages 12–22, Osaka, Japan, December 11-17 2016.

This knowledge is triggered by words during processing and affects the expectations on the upcoming input. Consequently, combinations that are new with respect to the already-stored knowledge require more cognitive efforts to be unified in the semantic memory. Such effect has been shown at the discourse level by the *Dependency Locality Theory* (Gibson, 2000), proving that the introduction of new discourse referents is a complexity parameter.

Hagoort (2013; 2016) has proposed **Memory, Unification and Control** (MUC) as a general model for sentence comprehension that aims at accounting for such balance between storage and computation. The Memory component of the model refers to the linguistic knowledge that is stored in long-term memory. This includes *unification-ready structures* corresponding to **constructions** (Goldberg, 2006) represented by sets of **constraints** pertaining to the various levels of linguistic representation (phonology, syntax, and semantics) for that construction. Each constraint specifies how a given construction can combine with other constructions at a particular level of linguistic representation, as well as the result of such unification.[4] The Unification component refers to the assembly of pieces stored in memory into larger structures, with contributions from context. Unification is a constraint-based process, which attempts to solve the constraints defining the constructions. Unification operations take place in parallel at all the representation levels. Therefore, syntax is not the only combinatorial component (cf. also Jackendoff (2002)): constructions are combined into larger structures also at the semantic and phonological levels.

In this paper, we present a computational model of semantic complexity in sentence processing, which is strongly inspired by MUC. Our model integrates various insights from current research in distributional semantics and recent psycholinguistic findings, which highlight the key role of knowledge about event structure and participants stored in semantic memory and activated during language processing (McRae et al., 1998; McRae et al., 2005; McRae and Matsuki, 2009; Bicknell et al., 2010; Matsuki et al., 2011). Moreover, recent experiments in EEG showed that the activation of the so-called literal' word meanings is only carried out when necessary (Rommers et al., 2013). Following such findings, some recent theoretical proposals argued that words do not really have meaning, they are rather *cues* to meaning: sentence comprehenders use them to make inferences about the event or the situation that the speaker wants to convey (Elman, 2009; Elman, 2011; Kuperberg and Jaeger, 2015; Kuperberg, 2016). In particular, our model relies on the following assumptions:

- long-term semantic memory stores **Generalized Event Knowledge** (**GEK**). GEK includes people's knowledge of typical participants and settings for events (McRae and Matsuki, 2009);

- at least a (substantial) part of GEK derives from our linguistic experience and can be modeled with distributional information extracted from large parsed corpora. In this paper, we only focus on this distributional subset of GEK, which we refer to as GEK_D;

- during sentence processing, lexical items (and constructions in general) activate portions of GEK_D, which are then unified to form a coherent representation of the event expressed by the sentence.

The aim of this research is to propose a novel distributional semantic framework to model online sentence comprehension. Our two-fold goal is i.) to build an incremental distributional representation of a sentence, and ii.) to associate a **compositional cost** to such a representation to model the complexity of semantic processing. In particular, we argue that semantic complexity depends on two factors: a.) the availability and salience of "ready-to-use" event information already stored in GEK_D and cued by lexical items, and b.) the cost of unifying activated GEK_D into a coherent semantic representation, with the latter depending on the mutual semantic congruence of the events participants. We thus predict that sentences containing highly familiar lexical combinations like (1a) (*musician* is in fact a familiar subject of *play*) are easier to process than sentences expressing novel ones like (1b). Moreover, the complexity of novel combinations depends on how easily they fit with stored event knowledge.

[4] From a neurolinguistic perspective, Pulvermuller et al. (2013) recently proposed a model encoding such set of constraints at the brain level. The activation of a construction is carried out by means of *discrete combinatorial neuronal assemblies* (DCNAs), which encode the combinatorial links between the different objects of the construction itself.

In the following sections, we will present a global distributional semantic complexity score combining event activation and unification costs. As a first evaluation of our framework, we will use the semantic complexity score in a difficulty estimation task on the Bicknell dataset (Bicknell et al., 2010).

2 Related work

Some of the previous works applying Distributional Semantic Models (henceforth DSMs) to sentence processing focused on the problem of computing a *semantic surprisal* index for the words of the sentence, on the basis of what Hale (2001) has proposed for syntax, and defined as the negative logarithm of the probability of a word given its previous linguistic context. The higher the surprisal of a word, the lower its predictability, and high surprisal values have been shown to correlate with an increase in processing difficulty (Frank et al., 2013; Smith and Levy, 2013). Mitchell et al. (2010) proposed a model to compute surprisal, based on the product of a trigram language model and of a semantic component, based in turn on the weighted dot product of the semantic vector of a target word and of a history vector, representing its prior context. The authors interpolated their model with the output of an incremental parser and they evaluated it on the task of predicting word reading times in a test set extracted from the Dundee Corpus (Kennedy et al., 2003). Their results showed that the semantic component improves the predictions, compared to models based only on syntactic information.

Building on the work of Mitchell et al. (2010) and Mitchell (2011), Sayeed et al. (2015) tested a similar model on a multimodal language corpus (the AMI Meeting corpus; see Carletta (2007)), being able to predict spoken word pronunciation duration.

A totally different perspective was adopted by Lenci (2011): starting from the method for thematic fit estimations that was introduced in Baroni and Lenci (2010), the author presented a compositional distributional model for reproducing the expectation update on the filler of the patient slot of a verb, depending on how the agent slot had been saturated (for example, if the agent of the verb *to check* is *journalist*, likely patients will be things that journalists typically check, such as *source, spelling* etc.). Lenci tried to model explicitly the process through which we modify our predictions on upcoming linguistic input on the basis of our event knowledge: the saturation of an argument slot imposes new semantic constraints on the other positions yet to be filled, with the consequence that entities typically co-occurring with the agent become more plausible for the situation described by the sentence.

Coming to related works in experimental psychology, Pynte et al. (2008; 2009) measured vector proximity between the content words of an eye-tracking corpus by means of Latent Semantic Analysis (Landauer et al., 1998) to show how inspection times of a target word are affected by its semantic relatedness with the adjacent words in the same sentence.

In a more recent contribution, Johns and Jones (2015) presented a DSM that assumes the storage and retrieval of linguistic experiences as the fundamental operations of sentence processing, following a long tradition of exemplar theories of memory starting with Hintzman (1986; 1988). Each sentence in their model is encoded as a vector obtained by summing its word random vectors with permutations to account for the word position in the sentence (see Sahlgren et al. (2008)). Vectors that are similar to the one of the currently processed sentence (the so-called *memory traces*) are activated and then are summed into an *expectation vector*. Finally, the expectation vector is used to make predictions about forthcoming words and to construct sentence meaning.

3 An incremental model of sentence comprehension

We model the comprehension of a sentence as an incremental process driven by the goal of constructing a coherent representation of the event the speaker intends to communicate with the sentence. We assume there is a data structure called **situation model** (SM) (Zwaan and Radvansky, 1998) that is incrementally updated in working memory during language comprehension. Given a sentence s being processed,[5] SM contains a representation of the event e_s described by s, which is compositionally built from GEK_D retrieved from long-term memory, and activated by the words in s. Similarly to MUC, our model is

[5] Although in this paper we focus on sentence processing, our model can equally apply at the sub-sentential level, such as phrases or chunks, as well as at the discourse level.

formed by a **memory component** containing lexical information, and a **unification component** dealing with the compositional construction of the sentence semantic representation.

4 The memory component: the representation of lexical knowledge

We assume the lexicon to be a repository of constructions (the latter including words along with more complex structures) stored in long-term memory. Each construction Cxn is defined by a form and a content. The latter consists of a set of pairs $\langle e_1, \sigma_1 \rangle, \ldots, \langle e_n, \sigma_n \rangle$, such that e_i is an event stored in GEK_D and σ_i is an **activation score**, expressing the salience of the event with respect to a construction and the strength with which the event is activated (cued) by the construction. At the moment, we assume that the activation score of the event *e* activated by Cxn is the conditional probability of the event given the construction, $P(e|Cxn)$.

We represent events in GEK with feature structures specifying their participants and roles, much like frames in Frame Semantics. More specifically, we represent the events in GEK_D, the distributional subset of GEK,[6] as feature structures containing information directly extracted from parsed sentences in corpora: attributes are **syntactic dependencies** (e.g. NSUBJ, NMOD-IN, etc.),[7] and values are **distributional vectors** of dependent lexemes.[8] The latter can be conceived as "out-of-context" distributional vector encoding of lexical items. Any type of distributional representation can be used to this purpose (e.g., explicit vectors, low-dimensionality dense embeddings, etc.). The following is a representation of an event $e \in GEK$, extracted from the sentence *The student reads the book.*:

(2) $[_{EVENT}$ NSUBJ:$\overrightarrow{student}$ HEAD:$\overrightarrow{read}$ DOBJ:$\overrightarrow{book}]$

Unlike previous syntax-based DSMs, we extract from corpora **syntactic joint contexts**, besides single dependencies (Chersoni et al., 2016). A syntactic joint context includes the whole set of dependencies of a given lexical head, which we assume as a surface representation of an event. Each event in GEK may be cued by several lexical items, as part of their semantic content, albeit with different strength depending on their statistical distribution. For instance, the event in (2) is cued by the noun *student*, the verb *read*, and the noun *book*.

We assume GEK to be hierarchically structured, according to various levels of event schematicity. In fact, all events in GEK can be *underspecified*. Without any need to add in GEK any specific structure, underspecification makes it possible to virtually generate **schematic events**, obtained by abstracting over one or more of its valued-attributes:

(3) a. $[_{EVENT}$ NSUBJ:$\overrightarrow{student}$ HEAD:$\overrightarrow{read}]$
b. $[_{EVENT}$ NSUBJ:$\overrightarrow{student}$ DOBJ:$\overrightarrow{book}]$

The feature structure in (3a) is a representation of a schematic event of a student reading, without any specification of the object, while (3b) represents an undespecified event involving a student acting on a book, which could be instantiated by specific events of reading, writing, buying, etc.

5 The unification component: constructing event representations

We assume that sentence comprehension always occurs within an existing SM and results into an update of this SM. The current SM acts as a constraint on the interpretation of the upcoming constructions, and it gets updated after the interpretation of every new construction. Sentence comprehension consists in recovering (reconstructing) the event e that the sentence is most likely to describe. The event e is the event that best satisfies all the **constraints** set by the constructions in the sentence and in the active SM. Let $w_1, w_2, \ldots, w_n$ be an input linguistic sequence (e.g., a sentence or a discourse) we have to interpret.

[6] In this paper we assume that $GEK = GEK_D$. Therefore, we henceforth omit the subscript for simplicity.

[7] We represent syntactic dependencies according to the Universal Dependencies annotation scheme (`http://universaldependencies.org/`).

[8] At this stage, we stay at the syntactic level, without entering into the mapping problem between syntactic and semantic arguments as described in (Dowty, 1991). All arguments in the event description correspond to syntactic roles, having in mind they could be used as a very rough approximation of semantic roles.

Let SM_i be the semantic representation built for the linguistic input until $w_1, \ldots, w_i$, and let e_i be the event representation in SM_i. When we process w_{i+1}:

i.) the GEK associated with w_{i+1} in the lexicon, $GEK_{w_{i+1}}$, is recovered;

ii.) $GEK_{w_{i+1}}$ is **integrated** with SM_i to produce SM_{i+1}, containing the new event e_{i+1}.

We model semantic composition as an **event construction and update function** F, whose aim is to build a coherent *SM* by integrating the GEK cued by the linguistic elements that are being composed:

$$F(SM_i, GEK_{w_{n+1}}) = SM_{i+1} \tag{1}$$

The composition function is responsible for two distinct processes:

1. F **unifies** two event feature structures into a new event. Given an event $e_i \in SM_i$ and $e_j \in GEK_{w_{n+1}}$, F produces a new event $e_k \in SM_{i+1}$:

$$F(e_i, e_j) = e_k = e_i \sqcup e_j \tag{2}$$

The unification function produces an output event if the attribute-values features of the input events are **compatible**, otherwise it fails. The following is an example of successful unification:

(4) $[_{EVENT}$ NSUBJ:$\overrightarrow{student}$ DOBJ:$\overrightarrow{thesis}] \sqcup [_{EVENT}$ NSUBJ:$\overrightarrow{student}$ HEAD:$\overrightarrow{read}] =$ $[_{EVENT}$ NSUBJ:$\overrightarrow{student}$ HEAD:$\overrightarrow{read}$ DOBJ:$\overrightarrow{thesis}]$

In this example, the event of a student acting on a thesis and the event of a student reading are unified into a new event of a student reading a thesis.

2. F **weights** the unified event e_k with a pair of scores $\langle \theta, \sigma \rangle$:

- θ is a score measuring the degree of **semantic coherence** of the unified event e_k. We assume that the semantic coherence (or internal unity) of an event depends on the **mutual typicality** of its components. Consider for instance the following sentences:

(5) a. The student reads a book.
b. The surfer reads a papyrus.

The event represented in (5a) has a high degree of internal coherence because all its components are mutually very typical: *student* is a typical subject for the verb *read* and *book* has a strong typicality both as an object of *read* and as an object related to *student*. Conversely, the components in the event expressed by (5b) have a low level of mutual typicality, thereby resulting into an event with much lower internal coherence. We model this idea of mutual typicality by extending the notion of **thematic fit**, which is normally used to measure the congruence of a predicate with an argument. In our case, instead, thematic fit is a general measure of the semantic typicality or congruence among the components of an event. In turn, we measure the thematic fit with vector cosine in the following way:

> Given s_i:a and s_j:b, such as s_i and s_j are two event attributes (e.g., NSUBJ, HEAD, etc.), the thematic fit of s_i:a with respect to s_j:b, $\theta(s_i{:}a, s_j{:}b)$, is the cosine between the vector of a and the prototype vector built out of the k most salient values $c_1, \ldots, c_k$, such that s_i:c_z, for $1 \leq z \leq k$, co-occurs with s_j:b in the same event structures.

For instance, the thematic fit of *student* as a subject of *read* is given by the cosine between the vector of *student* and the prototype vector built out of the k most salient subjects of *read*. Similarly, we also measure the typicality of *book* as an object related to *student* (i.e., the object of events involving student as subject) as the cosine between the vector of *book* and the prototype

vector built out of the k most salient objects related to *student*. Then we define the score θ of an event e as follows:

$$\theta_e = \prod_{a,b \in e} \theta(s_i{:}a, s_j{:}b) \tag{3}$$

Therefore, the semantic coherence of an event is given by the product of the mutual thematic fit between its components. The higher is the mutual typicality between the elements of an event, the higher is its internal semantic coherence.

- σ weights the **salience** of the unified event e_k by combining the weights of e_i and e_j into a new weight assigned to e_k. In this paper, we combine the σ weights with the logistic function:

$$F(\sigma_i, \sigma_j) = \sigma_k = \frac{1}{1 + e^{-(\sigma_i + \sigma_j)}} \tag{4}$$

The score σ of the unified event thus measures the strength with which it is activated (cued) by the composed linguistic expressions. This entails that events that are cued by more linguistic constructions in a sentence should incrementally increase their salience.

To sum up, we conceive composition as event unification. Unified events are weighted along two dimensions: internal semantic coherence (θ), and degree of activation by linguistic expressions (σ). These two dimensions also determine the **composition cost** of the unification process. We argue that the semantic complexity of a sentence s is inversely related to the sum of θ and σ:

$$SemComp_s = \frac{1}{\theta_s + \sigma_s} \tag{5}$$

The less internally coherent is the event represented by the sentence and the less strong is its activation by the lexical items, the more the unification is cognitively expensive and the sentence semantically complex. This is consistent with the MUC model of sentence comprehension: the harder is to build an integrated semantic representation through unification, the harder the processing effort, as reflected by a larger N400 amplitude.

6 Evaluation

As a first test for our framework, we measure the semantic complexity of the sentences in the Bicknell dataset (Bicknell et al., 2010). The Bicknell dataset was prepared to verify the hypothesis that the typicality of a verb direct object depends on the subject argument. For this purpose, the authors selected 50 verbs, each paired with two agent nouns that altered the scenario evoked by the subject-verb combination. Plausible patients for each agent-verb pair were obtained by means of production norms, in order to generate triples where the patient was *congruent* with the agent and with the verb. For each congruent triple, they also generated an *incongruent* triple, by combining each verb-congruent patient pair with the other agent noun, in order to have items describing atypical situations.

The final dataset included 100 pairs subject-verb-object triples, that were used to build the sentences for a self-paced reading and for an ERP experiment.[9] To give an example, experimental subjects were presented with sentence pairs such as:

(6) a. The journalist checked the spelling of his latest report. (*congruent condition*)
b. The mechanic checked the spelling of his latest report. (*incongruent condition*)

The sentences of each pair contain the same verb and the same object, differing for the subject. Given the subject, the object is a preferred argument of the verb in the congruent condition, whereas it is an implausible filler in the incongruent condition. Bicknell et al. (2010) reported that the congruent condition produced shorter reading times and smaller N400 signatures. Their conclusion was that verb argument expectations are dynamically updated during sentence processing, by integrating some kind of general

[9] Actually, Bicknell et al. (2010) used only a subset of 64 pairs, after removing the items that were potentially problematic for their experiments. In the present study, we use the original dataset.

knowledge about events and their typical participants. Lenci (2011) evaluated his model on the ability to assign a higher thematic fit score to the congruent triples than to the incongruent ones. We interpret Bicknell's experimental data as suggesting that congruent sentences are less semantically complex than incongruent sentences. Consistently, we predict that our model will assign a higher semantic complexity score to incongruent sentences than to congruent ones.

6.1 Modeling the GEK

Following the procedure described in Chersoni et al. (2016), we extracted from parsed corpora the syntactic joint contexts for all the words of the Bicknell triples. For our extraction, we used a concatenation of four different corpora: the British National Corpus (BNC; Leech (1992)); the Reuters Corpus vol.1 (RCV1; Lewis et al. (2004)); the ukWaC and the Wackypedia Corpus (Baroni et al., 2009).

For each sentence, we generated a joint context by extracting the verb and its direct dependencies. Our dependency relations of interest are subject (NSUBJ), direct object (DOBJ), indirect object (IOBJ) and a generic prepositional complement relation (PREPCOMP), on which we mapped the complements introduced by a preposition. We discarded all the modifiers and we just keep the nominal heads. Here is an example of extracted syntactic joint context: *athlete-n-nsubj___win-v-head___medal-n-dobj___at-p+olympics-n-prepcomp*. For each joint context, we also generated all its dependency subsets to obtain the underspecified schematic events. In total, we have extracted 4,204,940 syntactic joint contexts (including schematic events).

The collection of syntactic joint contexts were used to define the feature structures of the events in GEK, and cued by the target words of the Bicknell dataset. As described in section 4, each verb and noun occurring in these event structures was represented with a distributional vector in a syntax-based DSM using as contexts the dependencies extracted from the above corpora (e.g., `enemy-n:obj`).[10]

6.2 Computing the semantic complexity scores for the test sentences

The sentences in the original Bicknell dataset were first turned into S(subject)-V(erb)-O(bject) triples (e.g. NSUBJ:*journalist* HEAD:*check* DOBJ:*spelling*). For each test sentence s we computed σ_s and θ_s in the following way:

- σ_s We take the activation strength of the joint context formed by the test triple given S (i.e., σ_S), V (i.e., σ_V) and O (i.e., σ_O). For instance, σ_S is the activation strength of the joint context NSUBJ:*journalist* HEAD:*check* DOBJ:*spelling*, given *journalist*. Then σ_s is obtained by applying equation (4) to the sum of σ_S, σ_V and σ_O.

- θ_s This score represents the semantic coherence of the event represented by s and is obtained by measuring the mutual typicality of its components. Following equation (3), we compute θ_s as the product of the thematic fit of S given V, $\theta_{S,V}$, O given V, $\theta_{O,V}$, and the thematic fit of O given S, $\theta_{O,S}$. In particular, $\theta_{S,V}$ is the cosine between the vector of S and the centroid vector built out of the k most salient subjects of V (e.g., the cosine between the vector of *journalist* and the centroid vector of the most salient subjects of *check*), $\theta_{O,V}$ is the cosine between the vector of O and the centroid vector built out of the k most salient direct objects of V (e.g., the cosine between the vector of *spelling* and the centroid vector of the most salient objects of *check*), and $\theta_{O,S}$ is the cosine between the vector of O and the centroid vector built out of the k most salient direct objects occurring in events whose subject is S (e.g., the cosine between the vector of *spelling* and the prototype vector of the most salient objects of events whose subject is *journalist*). Following Baroni and Lenci (2010), we measured argument salience with LMI (Evert, 2005) and we fixed $k = 20$.

The final semantic complexity score $SemComp_s$ is the inverse of the sum of the σ and θ scores (see equation (5)). Notice that if the event corresponding to the sentence is not stored in GEK, its activation score is 0, and therefore the σ_s component will be null. In this case, the only relevant factor for semantic complexity is the event coherence measured by θ_s. This is consistent with the model we presented in

[10] We also use inverse dependencies (see Baroni and Lenci (2010)) in order to represent the relation of a target noun with its verb head: for example, given the sentence *The dog runs.*, the context of the target *dog-n* for this sentence will be *run-v:sbj-1*.

section 1 and based on the assumption that sentence processing is the result of a balance between retrieval of stored information and the building of new events through unification. If s describes a familiar event already stored in long-term memory as modelled with GEK, the complexity of s depends on how strong such event is cued by the lexical items in s and by the mutual typicality of its components. On the other hand, if the sentence describes a new event, its complexity only depends on the internal coherence of the event produced through unification.

7 Results and conclusions

For 16 pairs of triples of the Bicknell dataset we were not able to compute thematic fit scores, so we had to discard them.[11] Therefore, we are left with 84 pairs of triples (168 triples in total): in each triple, the patient is either typical (congruent) or atypical (incongruent) with respect to the agent.

First of all, the SemComp scores assigned to sentences in the congruent condition are significantly lower than the scores assigned to sentences in the incongruent conditions, according to the Wilcoxon test (W = 4791, p-value < 0.001). Our semantic complexity score is therefore able to model the higher processing difficulty of the incongruent sentences, as shown in the EEG experiments by Bicknell et al. (2010). We also evaluated the model accuracy, as the percentage of congruent sentences to which the model assigns a semantic complexity lower than score assigned to the incongruent sentence in the same pair. The model performance is compared with the random accuracy baseline, as in Lenci (2011).

Model	Hits	Accuracy	Significance
$\sigma_s + \theta_s$	62	73.8%	$p < .05$
θ_s	59	70.2%	$p < .05$
Baseline	42	50%	

Table 1: Number of hits and accuracy with or without σ scores. p-values computed with the χ^2 test.

Since the σ component is an element of novelty with respect to thematic fit-only models, we decided to test the algorithm also without it, that is to say to assign the complexity score only on the basis of the event semantic coherence. Although the difference is not huge, it is noteworthy that the σ component improves the accuracy score, supporting our hypothesis that semantic complexity depends both on retrieval and on unification costs.

Our model achieves exactly the same accuracy as the Expectation Composition and Update Model (ECU) in Lenci (2011) when evaluated on the same 84 triples (73.8%). However, it should be pointed out that ECU was tailored on the structure the Distributional Memory tensor (Baroni and Lenci, 2010) and on the Bicknell dataset. Indeed, the ranking function for the typical fillers of a slot depends on the availability in the tensor of syntactic relations (in this case, the OBJ and the VERB relation) that can be used as simultaneous constraints on a candidate. In other words, given a patient *p*, an agent *AG* and verb *v*, *p* has to have a high association score both in the triple {*p*, OBJ, *v*} and in the triple {*AG*, VERB, *p*}. These relations work well for representing constraints on the agent and on the patient slot, but it is not clear how ECU could estimate expectations on other slots, say the instrument and/or the location one. Moreover, it does not take into account the Memory component.

Our results are obtained with a much more general model of semantic complexity that can be applied to any type of syntactic structure (the set of syntactic relations that we consider in the extraction of the joint contexts is a parameter) and is based on a less *ad hoc* and more sophisticated distributional representation of GEK. Concerning the σ component, we should also mention that a joint context for the full event was retrieved for only 22 of the 168 triples. As expected, an implementation of the memory component based only on textual corpora suffer from data sparsity, and the future developments of this model will have to take this factor into account. The introduction of a robust generalization component, which could generate new joint contexts by making inferences on new potential event participants, could help to mitigate such problem.

[11] We discarded from the syntax-based DSM words with a frequency below 100 in the training corpus. Consequently, for some triples one or more words did not have any vector representation in the DSM, so that we could not compute the thematic fit scores that are required by our model.

The semantic complexity model we have proposed in this paper is strongly inspired by the general cognitive principles of the MUC architecture. In particular, we rely on two components to assign semantic complexity scores: i) a memory component, consisting of a distributional subset of GEK, such that the more an event is strongly activated by linguistic cues, the easier will be its retrieval from the semantic memory; ii) a unification component, consisting of a composition and update function which unifies the GEK activated by linguistic cues into new structures. The more the unified components are mutually typical, the more semantically coherent will be the event. Our assumption is that linguistic constructions that are strongly activated by the previous context and with high values of semantic coherence are easier to process. In the future, we plan to extend our experimentations to a wider range of psycholinguistic datasets, in order to see how the model can deal with a larger number of complexity sources and linguistic structures.

Hopefully, future extensions of this model will also present a more global notion of complexity and will integrate information coming from different linguistic domains. It would be interesting, for example, to combine the predictions of our model of semantic complexity with constraint-based frameworks for the estimation of syntactic difficulty, such as Blache's Property Grammars (Blache, 2011; Blache, 2013), and to see how they correlate with experimental data.

There are many other aspects in language processing that, at the moment, our model leaves aside. Future extensions, in our view, should also account for the role played by attention,[12] since several linguistic devices (prosodic cues, non-canonical syntactic structures etc.) can be used to signal informationally relevant parts of the message to the listeners/readers, helping them in the allocation of processing resources and thus influencing complexity (Hagoort, 2016). At the best of our knowledge, such issues have still to be convincingly addressed by current models.

8 Acknowledgements

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

References

Giosuè Baggio and Peter Hagoort. 2011. The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes*, volume 26(9): 1338-1367.

Giosuè Baggio, Michiel van Lambalgen, and Peter Hagoort. 2012. The processing consequences of compositionality. *The Oxford Handbook of Compositionality*, Oxford University Press, Oxford, 1-23.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, volume 43(3): 209-226.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, volume 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. *Journal of Memory and Language*, volume 63(4): 489-505.

Philippe Blache. 2011. Evaluating Language Complexity in context: new parameters for a constraint-based model. *Proceedings of CSLP-2011*.

Philippe Blache. 2013. Chunks et activation: un modèle de facilitation du traitement linguistique. *Proceedings of TALN-2013*.

Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, volume 41(2): 181-190.

[12] We thank one of our anonymous reviewers for pointing this out. On a related topic, we would like to refer the readers to Zarcone et al. (2016) for a systematic overview on the use of the notions of salience and attention in surprisal-based models of language processing.

Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2016. Representing verbs with rich contexts: an evaluation on verb similarity. *Proceedings of EMNLP*.

Noam Chomsky. 1957. Syntactic Structures. Mouton & Co.

David Dowty. 1991. Thematic proto-roles and argument selection. Language, 67(3):547-619.

Jeffrey L Elman. 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, volume 33(4): 547-582.

Jeffrey L Elman. 2011. Lexical knowledge without a lexicon? *The Mental Lexicon*, volume 6(1): 1-34.

Jeffrey L Elman. 2014. Systematicity in the lexicon: on having your cake and eating it too. *The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*, edited by Paco Calvo and John Symons, The MIT Press, Cambridge, MA.

Stefan Evert. 2005. The statistics of word cooccurrences: word pairs and collocations.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2013. Word surprisal predicts N400 amplitude during reading. *Proceedings of ACL*: 878-883.

Ted Gibson. 2000. The Dependency Locality Theory: a Distance-Dased Theory of Linguistic Complexity. *Image, Language, Brain*, edited by Alec Marantz, Yasushi Miyashita and Wayne O'Neil, MIT Press.

Adele E Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Language*, Oxford, Oxford University Press.

Peter Hagoort. 2003. Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *Journal of Cognitive Neuroscience*, volume 15(6).

Peter Hagoort. 2013. MUC (memory, unification, control) and beyond *Frontiers in Psychology*, volume 4: 1-13.

Peter Hagoort. 2016. MUC (Memory, Unification, Control): A Model on the Neurobiology of Language Beyond Single Word Processing In G. Hickok and S. Small (eds.), *Neurobiology of Language*, Amsterdam, Elsevier: 339-347.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. *Proceedings of NAACL-HLT*: 1-8.

Douglas L Hintzman. 1986. 'Schema abstraction' in a multiple-trace memory model. *Psychological Review*, volume 93(4): 411-428.

Douglas L Hintzman. 1988. Judgements of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, volume 95(4): 528-551.

Ray Jackendoff. 2002. Foundations of Language: Brain, Meaning, Grammar, Evolution. Cambridge, Cambridge University Press.

Alan Kennedy, Robin Hill, and Joel Pynte. 2003. The Dundee Corpus. *Proceedings of the 12th European Conference on Eye Movement*.

Gina R Kuperberg and Florian T Jaeger. 2015. What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, volume 31(1): 32-59.

Gina R Kuperberg. 2016. Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, volume 31(5): 602-616.

Marta Kutas and Kara D Federmaier. 2000. Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, volume 4(12): 463-470.

Brendan T Johns and Michael Jones. 2015. Generating structure from experience: a retrieval-based model of language processing. *Canadian Journal of Experimental Psychology*, volume 69(2).

Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, volume 25(2-3): 259-284.

Geoffrey Leech. 1992. 100 million words of English: the British National Corpus (BNC). *Language research*, volume 28(1): 1-13.

Alessandro Lenci. 2011. Composing and updating verb argument expectations: a distributional semantic model. *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*: 58-66.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv 1: A new benchmark collection for text categorization research. *The journal of Machine Learning research*, volume 5: 361-397.

Kazunaga Matsuki, Tracy Chow, Mary Hare, Jeffrey L Elman, Christoph Scheepers, and Ken McRae. 2011. Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, volume 37(4): 913-934.

James L McClelland. 1994. The interaction of nature and nurture in development: A parallel distributed processing perspective. *International perspectives on psychological science*, volume 1: 57-88.

Ken McRae, Michael J Spivey-Knowlton and Michael K Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, volume 38(3): 283-312.

Ken McRae, Mary Hare, Jeffrey L Elman, and Todd Ferretti. 2005. A basis for generating expectancies for verbs from nouns. *Memory and Cognition*, volume 33(7): 1174-1184.

Ken McRae and Kazunaga Matsuki. 2009. People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and linguistics compass*, volume 3(6): 1417-1429.

Laura A Michaelis. 2013. Sign-Based Construction Grammar. *The Oxford Handbook of Construction Grammar*, edited by Thomas Hoffmann and Graeme Trousdale, Oxford University Press, Oxford: 133-152.

Jeff Mitchell, Mirella Lapata, Vera Demberg and Frank Keller. 2010. Syntactic and semantic factors in processing difficulty: An integrated measure. *Proceedings of ACL*: 196-206.

Jeff Mitchell. 2011. Composition in distributional models of semantics. PhD Thesis, The University of Edinburgh.

Friedemann Pulvermuller, Bert Cappelle and Yury Shtyrov. 2013. Brain basis of meaning, words, constructions, and grammar. *The Oxford Handbook of Construction Grammar*, edited by Thomas Hoffmann and Graeme Trousdale, Oxford University Press.

Joel Pynte, Boris New, and Alan Kennedy. 2008. On-line contextual influences during reading normal text: A multiple-regression analysis. *Vision research*, volume 48(21): 2172-2183.

Joel Pynte, Boris New, and Alan Kennedy. 2009. On-line contextual influences during reading normal text: The role of nouns, verbs and adjectives. *Vision research*, volume 49(5): 544-552.

Milena Rabovsky and Ken McRae. 2014. Simulating the N400 ERP component as semantic network error: insights from a feature-based connectionist attractor model of word meaning. *Cognition*, volume 132(1): 68-89.

Joost Rommers, Ton Dijkstra and Marcel Bastiaansen. 2013 Context-dependent Semantic Processing in the Human Brain: Evidence from Idiom Comprehension. *Journal of Cognitive Neuroscience*, 25(5):762-776.

Ivan A Sag. 2012. Sign-Based Construction Grammar: An Informal Synopsis. *Sign-Based Construction Grammar*, edited by Hans C Boas and Ivan A Sag, CSLI Publications, Stanford, CA: 61-197.

Magnus Sahlgren, Anders Holst, and Pentti Kanerva. 2008. Permutations as a mean to encode order in word space. *Proceedings of the 30th Conference of the Cognitive Science Society*: 1300-1305.

Asad Sayeed, Stefan Fischer and Vera Demberg. 2015. Vector-space calculation of semantic surprisal for predicting word pronunciation duration. *Proceedings of ACL*: 763-773.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, volume 128(3): 302-319.

Jos JA Van Berkum, Colin M Brown, Pienie Zwitserlood, Valesca Kooijman, and Peter Hagoort. 2005. Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, volume 31(3): 443-467.

Alessandra Zarcone, Marten Van Schijndel, Jorrig Vogels and Vera Demberg. 2016. Salience and attention in surprisal-based accounts of language processing. *Frontiers in Psychology*, volume 7.

Rolf A Zwaan and Gabriel A Radvansky. 1998. Situation models in language comprehension. *Psychological Bulletin*, volume 123(2): 162-185.