Language Resources and Evaluation (2021) 55:873–900
https://doi.org/10.1007/s10579-021-09533-9

ORIGINAL PAPER

Not all arguments are processed equally: a distributional model of argument complexity

Emmanuele Chersoni[1] · Enrico Santus[2] · Alessandro Lenci[3] · Philippe Blache[4] · Chu-Ren Huang[1]

Accepted: 25 January 2021 / Published online: 3 March 2021

Abstract

This work addresses some questions about language processing: what does it mean that natural language sentences are semantically complex? What semantic features can determine different degrees of difficulty for human comprehenders? Our goal is to introduce a framework for argument semantic complexity, in which the processing difficulty depends on the typicality of the arguments in the sentence, that is, their degree of compatibility with the selectional constraints of the predicate. We postulate that complexity depends on the difficulty of building a semantic representation of the event or the situation conveyed by a sentence. This representation can be either retrieved directly from the semantic memory or built dynamically by solving the constraints included in the stored representations. To support this postulation, we built a Distributional Semantic Model to compute a compositional cost function for the sentence unification process. Our evaluation on psycholinguistic datasets reveals that the model is able to account for semantic phenomena such as the context-sensitive update of argument expectations and the processing of logical metonymies.

✉ Emmanuele Chersoni
emmanuele.chersoni@polyu.edu.hk

Enrico Santus
esantus@mit.edu

Alessandro Lenci
alessandro.lenci@unipi.it

Philippe Blache
philippe.blache@univ-amu.fr

Chu-Ren Huang
churen.huang@polyu.edu.hk

1 Hong Kong Polytechnic University, Hung Hom, Hong Kong

2 Massachusetts Institute of Technology, Cambridge, USA

3 University of Pisa, Pisa, Italy

4 Aix-Marseille University, Marseille, France

Keywords Distributional semantics · Psycholinguistics · Argument complexity · Logical metonymy · Cognitive modeling

1 Argument semantic complexity

It is a well-known fact that nouns differ for their acceptability as predicate arguments. Traditionally, linguistic theory has modeled this as a binary contrast between acceptable vs. impossible arguments:

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

Impossible arguments are those that violate the combinatorial constraints (aka **selectional preferences**) of the predicate to such a degree that we are not able to build any coherent representation for the described situation, like in (1-b). Recently, psycholinguistic and neurocognitive research has questioned the dichotomous nature of the phenomenon, arguing that arguments differ in their degree of acceptability, as shown by the following sentence:

(2) The gardener plays the castanets in the cave.

Although the selectional constraints of *play* are satisfied both in (1-a) and (2), the latter expresses a more unusual event. Investigations on event-related potentials (ERP)—the electrophysiological responses of the brain to a stimulus measured with electroencephalography (EEG)—have brought extensive evidence that sentences like (1-a) and (2), despite being both semantically acceptable, have a different cognitive status. In particular, sentences such as (2), including possible but unexpected combinations of lexemes, evoke stronger N400 components than plausible ones. The N400 component, originally described by Kutas and Hillyard (1980), is a negative-going deflection that peaks around 400 ms after the presentation of the stimulus, and since its discovery the amplitude of this component has been taken to reflect *the complexity of semantic composition*: unusual combinations of lexemes require an extra cognitive effort to be understood, as they are not coherent with the unfolding semantic representation of the context (Baggio and Hagoort 2011; Baggio et al. 2012). We refer to this phenomenon as **argument complexity**, to distinguish it from other cases of syntactic and semantic complexities occurring during online sentence comprehension. Over the years, several linguistic theories and computational models have been proposed to account for processing differences between natural language sentences, among which we can cite Dependency Locality Theory (Gibson 2000), the ACT-R based model by Lewis and Vasishth (2005) and Surprisal Theory (Hale 2001, 2016). A common point of all the above-mentioned frameworks is the focus on the syntactic factors of complexity, which are sometimes identified with the length of dependencies, sometimes with the probabilities of a given syntactic analysis in a given context, and so on. The notion of argument complexity we analyse in this work instead concerns the semantic factors leading to the construction of sentence meaning via predicate-argument composition.

A different but strongly related phenomenon is **(complement) coercion**, in which an argument is reinterpreted to overcome the violation of its predicate selectional preferences (Lauwers and Willems 2011). One widely studied case of coercion is **logical metonymy**, which is traditionally considered as a theoretical challenge for classical models of compositionality (Pustejovsky and Batiukova 2019):

(3) The author began the book.

Logical metonymy is described as a type clash between an event-selecting metonymic verb (e.g., aspectual verbs like *begin*) and an entity-denoting nominal object, which triggers the recovery of a hidden event (e.g., *writing*). Crucially, previous research has brought extensive evidence that such metonymic constructions also determine extra processing costs and increased complexity during online sentence comprehension (McElree et al. 2001; Traxler et al. 2002), apparently due to "the deployment of operations to construct a semantic representation of the event" (Frisson and McElree 2008). Therefore, logical metonymy, as well as complement coercion in general, can be regarded as an instance of argument complexity caused by the effort required to repair the violation of the verb selectional preferences.

The N400 effects and the processing costs of logical metonymy suggest that "not all arguments are processed equal", and that the semantic complexity of an argument depends on its compatibility with the selectional constraints of the predicate. Argument compatibility is a graded, rather than binary notion and is typically referred to as **thematic fit**. Several psycholinguistic studies making use of different experimental paradigms (self-paced reading, eye-tracking, EEG, etc.) indicate that argument complexity is determined by information about event contingencies and specific predicate-argument combinations stored in semantic memory. This event knowledge has a key role in human sentence processing: Verbs (e.g., *eat*) activate expectations about nouns typically occurring as their arguments (e.g., *pizza*) (McRae et al. 1998), and in turn entity-denoting nouns prime verbs referring to the events they typically participate in (McRae et al. 2005). Arguments that are coherent with the activated expectations have a lower semantic complexity and are read faster by subjects.

Moreover, priming experiments show that nouns trigger also other nouns co-occurring as arguments in the same events (Hare et al. 2009). More in detail: (i) nouns of events prime participants (*sale-shopper*) and objects (*trip-luggage*) typically found at those events; (ii) locations prime people/animals and objects (*hospital-doctor, barn-hay*) typically found at those locations; (iii) instrument nouns prime things on which they are commonly used (*key-door*). All these event-based priming effects support the hypothesis of a mental lexicon arranged as *a web of mutual expectations* that are exploited online to compute the thematic fit of the argument nouns as fillers of the verb roles. In the literature, this knowledge contained in semantic memory is generally referred to as *Generalized Event Knowledge* (*GEK*), and it is acquired by humans from first-hand experience (e.g., playing music) and linguistic experience (e.g., talking and reading about playing music) (McRae and Matsuki 2009).

The expectations for the predicate fillers and the resulting argument complexity depend on wide event scenarios. As shown by some recent studies (Bicknell et al. 2010; Matsuki et al. 2011), the expectations about the likely filler of a given verb

argument (e.g., the patient role) depend on the fillers of the other verb arguments (e.g., the agent). For example, given an agent noun like *boxer*, the most likely patient for the verb *dodge* is *punch*, while if the agent noun is *politician*, something like *question* will be much more likely as a patient filler. In other words, argument complexity and thematic fit have a context-sensitive nature and are affected by the general situation described by the sentence. Sentences including congruent argument combinations elicit significantly smaller N400 amplitudes than incongruent ones (Bicknell et al. 2010), as they show lower processing complexity. After an analysis of the evidence presented in the previously-cited studies, Jeffrey Elman proposed that words should be conceived as cues to event knowledge (*words-as-cues hypothesis*), and that sentence meaning consists precisely of the event representations that the lexical items in the sentence activate (Elman 2009, 2014). As new information comes in during online linguistic processing, new constraints on the possible interpretations of the sentence are progressively added. Importantly, logical metonymy too is affected by the whole configuration of verb arguments. For instance, the event recovered to overcome the type clash depends on both the patient and the agent roles (Lascarides and Copestake 1998; Zarcone et al. 2014). Therefore, argument complexity in general is a compositional phenomenon that must be addressed within the context of the cognitive processes leading to sentence meaning construction.

1.1 Argument complexity in distributional semantics

Computational models of argument complexity have been extensively investigated in distributional semantics (Lenci 2018). Erk et al. (2010) were among the first authors to measure the correlation between human-elicited thematic fit ratings and the scores assigned by a syntax-based Distributional Semantic Model (DSM). The plausibility of each verb-argument pair was computed as the similarity between new candidate nouns and previously attested exemplars for each specific verb-role pairing, as already proposed in Erk (2007). Baroni and Lenci (2010) adopted an approach to thematic fit modeling that has become dominant in the literature: For each verb role, they used their Distributional Memory (henceforth DM) framework to build a 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. Lenci (2011) later extended this model to account for the dynamic update of the expectations on an argument, depending on how another role is filled. By using the same DM tensor, this study tested an additive and a multiplicative model (Mitchell and Lapata 2010) to compose and update the expectations on the patient filler of the subject–verb–object triples of the dataset used in the study by Bicknell et al. (2010). More recent contributions aimed at improving the original model by Baroni and Lenci (2010), either by using semantic role labels instead of syntactic dependencies as the context for the vectors (Sayeed et al. 2015) or by clustering the verb fillers in order to better deal with polysemy (Greenberg et al. 2015). Another variant of the model, introduced by Santus et al. (2017), achieves better results by replacing cosine with a metric based on the semantic feature overlap between the prototype and the candidate fillers.

A different approach to the thematic fit problem was proposed by Tilk et al. (2016), who presented two neural architectures for generating probability distributions over the possible arguments for each thematic role. Their models took advantage of supervised training on two role-labeled corpora to optimize the distributional representation for thematic fit modeling, and managed to obtain significant improvements over the other systems on almost all the evaluation datasets. They also evaluated their model on the task of composing and updating verb argument expectations, obtaining a performance comparable to Lenci (2011). More recently, Chersoni et al. (2019) proposed a general distributional model for incremental sentence meaning representation that has been tested on human ratings of compositional argument plausibility. A closely related notion to thematic fit is the one of selectional preference (Resnik 1997), with the main difference being that the former refers to a gradient compatibility between arguments and thematic roles, while the latter involves discrete semantic types (Lebani and Lenci 2018). The acquisition of selectional preferences has mostly been seen as an auxiliary task for improving the performance of systems with different goals, such as semantic role classification (Collobert et al. 2011; Zapirain et al. 2013; Roth and Lapata 2015) or coreference resolution (Heinzerling et al. 2017). Some recent and some notable exceptions are the studies by Zhang et al. (2019, 2020), which introduced large-scale evaluation benchmarks for the task and proposed multiplex embedding models incorporating both the overall semantics of a word and its relational dependencies in context.

Concerning the Natural Language Processing (NLP) research on logical metonymy, previous studies focused on two different and complementary aspects of the phenomenon. On the one hand, the retrieval of the covert event, which has been approached by means of either probabilistic methods (Lapata and Lascarides 2003) or distributional similarity models (Zarcone et al. 2012). On the other hand, the modeling and reproduction of the processing differences observed in the experimental literature, a problem mainly tackled, again, with DSMs (Zarcone et al. 2013). In our view a computational model, in order to provide a complete account of logical metonymy and its processing consequences, should be able to deal with both of these aspects.

Leveraging and extending these previous results, we introduce in Sect. 2 a distributional model of argument complexity inspired by the Memory, Unification and Control framework by Hagoort (2013). Our proposal has two major elements of novelty. First of all, it is able to subsume the gradient nature of argument acceptability in (2) and the coercion in (3) under the same general computational approach to argument complexity. Secondly, it is grounded on the assumption that distributional semantics can provide a useful model of (at least a subset of) GEK and of its role in constructing compositional semantic representations. In Sect. 3, we evaluate our model on two psycholinguistic datasets, respectively, in the task of composing and updating verb argument expectations and in modeling logical metonymy.

2 A distributional model for argument complexity

The objectives of our model are (i) to build an incremental distributional representation of a sentence, and (ii) to introduce a **compositional weight** to account for its

argument complexity. We assume that sentences represent events consisting of various participants playing different roles, and that their argument complexity depends on two main factors: (a) the availability and *salience* of "ready-to-use" event information already stored in GEK and (b) the cost of *unifying* the GEK portions activated by the context into a coherent semantic representation, a cost mainly depending on the mutual *semantic coherence* of the event participants. We thus predict that sentences containing highly *familiar argument combinations* are easier to process than sentences containing novel ones, like the one in (2). Moreover, the complexity of novel combinations depends on how "compatible" they are with the event knowledge stored in the semantic memory.

GEK is assumed to be a highly structured repository, organized under various levels of complexity, granularity, and schematicity. It includes information about fully-specified micro-events (e.g., students reading books, gardeners cutting grass, etc.) and about more complex scenarios. In fact, sentences can be regarded as *partial descriptions of events*, since many details about described situations can be left unspecified, and it is up to the comprehender to infer the missing parts by retrieving relevant information in GEK: for example, when we hear a sentence like *The soldier killed all the enemies*, we could infer that he used some sort of weapon (e.g., a rifle, a machine gun, etc.) to perform the killing event. Consistently with the psycholinguistic findings reviewed in Sect. 1 and with Elman's words-as-cues hypothesis, each linguistic expression works as a cue for recovering portions of GEK. Not only verbs, but also nouns (and possibly adjectives) activate GEK: more specifically, they activate the events involving those entities. For instance, hearing the noun *student* in a sentence leads to the activation of *student*-related events in GEK. As long as comprehenders manage to retrieve the "right" event scenario, they are also able to anticipate upcoming arguments in the sentence, and fill in unexpressed elements (e.g, like the covert event in logical metonymic sentences).

Comprehension consists in recovering the most likely event expressed by a sentence (Kuperberg 2016), and it is an incremental process leading to the construction of a semantic representation, which is in turn obtained by combining the subsets of GEK activated by the different constructions in the sentence. Analogously to Hagoort (2013), we distinguish between two components of our model:

- a **Memory** component, representing the storage of event structures in GEK contained in the semantic memory. In this study, we only consider the GEK subset derived from linguistic experience, which we model with distributional information extracted from corpora;
- a **Unification** component, which combines the GEK portions activated by linguistic expressions, in order to generate new and more complex structures.

2.1 The memory component: modeling *GEK* in long-term memory

In our framework, we assume that each lexical item w_i activates a set of events $\langle e_1, \sigma_1 \rangle, \ldots, \langle e_n, \sigma_n \rangle$ such that e_i is an event in GEK, and σ_i is an activation score computed as the conditional probability $P(e|w_i)$. In other words, the activation level of e is quantified as its probability given the context word w_i. Therefore, processing

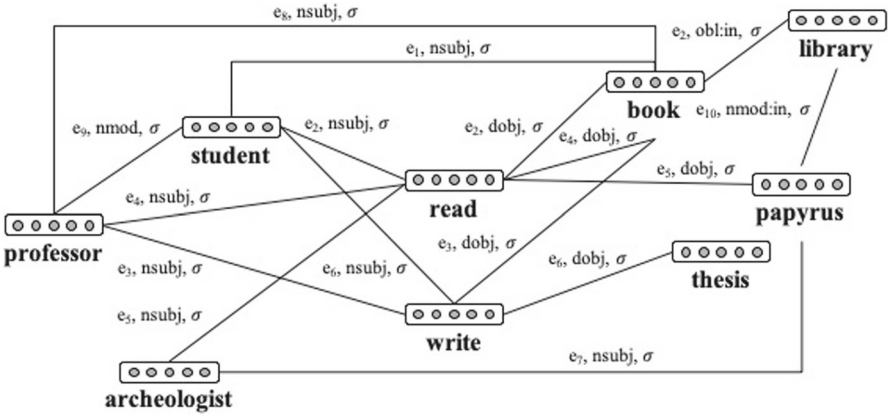

Fig. 1 A fragment of the DEG representing GEK with several instances of events, each represented by a sequence of co-indexed e. The σ are the activation scores of events

a linguistic expression in a given sentence will lead to the activation of a set of events in the semantic memory, each one associated to a σ score.

In a previous work, Chersoni et al. (2019) represented GEK with a *Distributional Event Graph* (DEG) that contains events extracted from dependency parsed sentences (Fig. 1).[1] The DEG nodes were distributional vectors (i.e., embeddings), meant as "out-of-context" encoding of lexical items. Notice that, in principle, any type of distributional vector can be used to this purpose. The edges corresponded to syntactic relations as an approximation of deeper semantic roles (e.g., the subject relation for the agent, the direct object relation for the patient, etc.), and they were weighted with activation scores identifying the most prototypical event-entity links.

The approach that we followed for representing events, in this work, is to extract **syntactic joint contexts** (Chersoni et al. 2016b). A syntactic joint context includes the whole set of dependencies of a given lexical head (ignoring determiners and modifiers), and we assume it as a surface representation of an event. For example, from the dependency structure of the sentence *The student reads a book* we extract the following event corresponding to a path in the DEG in Fig. 1:

(4) $[_{E_2}$ NSUBJ:**student** HEAD:**read** DOBJ:**book**$]$

Events in GEK can be cued by several lexical items, with a strength depending on the salience of the event given the item. For example, the event above is cued by *student*, *read* and *book*. Besides complete events, we assume GEK to contain schematic (i.e., underspecified) events too, obtained by abstracting away from one or more arguments. For instance, from the sentence *The student reads a book* we also generate the schematic event $[_{E_1}$ NSUBJ:**student** DOBJ:**book**$]$ describing an underspecified event schema with a student agent and a book patient, which can be instantiated by different

[1] We represent dependencies according to the Universal Dependencies annotation scheme: `http://universaldependencies.org/`.

actions (e.g., *reading*, *borrowing*, etc.). Therefore, GEK is not a flat list of events, but a structured repository of prototypical knowledge about event contingencies.

2.2 The unification component: building semantic representations

Language can be seen as a set of instructions that the comprehender uses to represent the situation described by the speaker. In our framework, the event currently being processed is stored in a data structure called **Semantic Representation** (henceforth SR), which is akin to Discourse Representation Structures in DRT (Kamp 2013; Chersoni et al. 2019). Comprehension always occurs within the context of an existing SR: during online sentence processing, lexical items cue portions of GEK and the SR is dynamically updated by unifying the current content with the new information.

We anticipated that, in our view, the goal of 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 lexical items in the sentence and by the active SR. Let $w_1, w_2, \ldots, w_n$ be an input linguistic sequence (e.g., a sentence) that is currently being processed. Let SR_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 SR_i. When we process w_{i+1}:

1. $GEK[w_{i+1}]$, the event knowledge associated with w_{i+1} in the lexicon, is activated;
2. $GEK[w_{i+1}]$ is integrated with SR_i to produce SR_{i+1}.

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

$$F(SR_i, GEK[w_{i+1}]) = SR_{i+1}. \tag{1}$$

The composition function carries out two distinct processes:

- F **unifies** the events activated by two lexical items into a new complex event:

  (5) $[_{E_i}$ NSUBJ:**mechanic** DOBJ:**engine**$] \sqcup [_{E_j}$ NSUBJ:**mechanic** HEAD:**check**$] =$
  $[_{E_k}$ NSUBJ:**mechanic** HEAD:**check** DOBJ:**engine**$]$

  In this example, the event of a mechanic performing an action on an engine activated by the noun *mechanic* and the event of a mechanic checking something activated by the verb *check* are unified into a new complex event of a mechanic checking an engine;
- F **weights** the unified event e_k with a pair of scores $\langle \theta_{e_k}, \sigma_{e_k} \rangle$, weighting e_k with respect to its *semantic coherence* θ_{e_k} and to the *salience* σ_{e_k} of its activation.

Semantic coherence and activation salience, which will be illustrated in the following section, are the essential factors of our model of the argument complexity of semantic representations.

2.2.1 The cost of unification: semantic coherence

We introduce the score θ_{e_k} to quantify the degree of **semantic coherence** of a unified event e_k, under the assumption that such coherence depends on the **mutual typicality** of its components. Consider the following sentences:

(6) a. The student writes a thesis.
b. The mechanic writes a sonnet.

The event represented in (6-a) has a high degree of semantic coherence because all its components are mutually typical: *student* is a typical subject for the verb *write* and *thesis* has a strong typicality both as an object of *write* and as an object occurring in *student*-related events. Conversely, the components in the event expressed by (6-b) have a low level of mutual typicality, thereby resulting into an event with much lower semantic coherence. Although the sentence is perfectly understandable, the described situation is more unusual.

Verb-argument typicality is measured in the computational and psycholinguistic literature with *thematic fit* values (McRae et al. 1998). In the present proposal, the notion of thematic fit is extended in order to account for the degree of coherence of the events described by whole sentences. In computational approaches (Baroni and Lenci 2010), thematic fit is modeled with vector cosine in the following way:

> Given a list of lexemes $\mathbf{c_1}, \ldots, \mathbf{c_n}$ occurring in the same event structures as $\mathbf{b}$ with the role s_i and ordered by their decreasing salience, $\theta(\mathbf{a}|s_i, \mathbf{b})$ (the thematic fit of $\mathbf{a}$ given $\mathbf{b}$ and the role s_i) is the cosine between $\mathbf{a}$ and the prototype vector built out of the k top values $\mathbf{c_1}, \ldots, \mathbf{c_k}$, with for $1 \leq k \leq n$.

For instance, the thematic fit of *student* as an agent in *writing*-events is given by the cosine between the embedding of *student* and the centroid vector built out of the k most salient agents of *write*. Similarly, the typicality of *thesis* as a patient related to *student* (i.e., as a patient in events involving student as an agent) could be assessed by measuring the cosine between the embedding of *thesis* and the centroid vector built out of the k most salient patients related to *student*, and the typicality of *thesis* as a patient of *write* can be measured in the same way. In other words, typical fillers of a given role are used to build a sort of abstract distributional representation of an "ideal" filler for that role, and the thematic fit of a new candidate is computed as the distance between its embedding and the vector of the ideal filler.

Although we adopt the same approach for measuring the typicality of the participants, an important problem is *how the partial scores of single event-participant combinations are combined in a global semantic coherence score*. In our work, we experimented with two different solutions:

– as in Chersoni et al. (2016a) and Chersoni et al. (2017a), semantic coherence is assessed as *the product of all the partial thematic fit scores for all the event-participant (and inter-participant) combinations within a sentence*;[2]

[2] Beyond traditional calculations of thematic fit for the fillers of verb roles, we also compute scores for a generic co-participant relation between filler nouns, as experimental studies report processing facilitations

– similarly to Lenci (2011) and Chersoni et al. (2017b), semantic coherence is assessed as *the cosine similarity between the arguments of the sentence and the prototype vector of current argument expectations*, which is dynamically updated as new information from newly-saturated arguments comes in.

In the first case, the global score θ_{e_k} of an event e_k is defined as:

$$\theta_{e_k} = \prod_{a,b,s_i \in e} \theta(\mathbf{a}|s_i, \mathbf{b}) \tag{2}$$

For example, given a sentence like *The student drinks beer*, the score θ_{e_k} would be the product of three factors: (i) the thematic fit of *student* as an *agent* (AG) of *drink*; (ii) the thematic fit of *beer* as a *co-participant* (CO) of *student*; (iii) the thematic fit of *beer* as a *patient* (PA) of *drink*. Thus, θ_{e_k} would be computed as:

$$\theta_{e_k} = \theta(\mathbf{student}|AG, \mathbf{drink}) \cdot \theta(\mathbf{beer}|CO, \mathbf{student}) \cdot \theta(\mathbf{beer}|PA, \mathbf{drink}). \tag{3}$$

The product between thematic fit scores directly captures the idea of the mutual typicality between all event participants. Indeed, as an effect of the product, if the partial thematic fit score between an argument pair is low (e.g., the agent–patient combination), this will decrease the semantic coherence of the entire event. In the experiments in Sect. 3, we refer to the models using this computation of semantic coherence score as **ThetaProd** models.

The alternative approach consists in building a prototype vector for the final argument that needs to be predicted (e.g., the patient in an agent–verb–patient triple) using a single representation that incorporates the updated expectations for the verb given the previously-realized arguments (Lenci 2011; Chersoni et al. 2017b). In this model, the update on the expectation EX for a given filler caused by a new input word (e.g., a verb combining with an agent) is modeled with a function $f(x)$ that combines the prototypes built out of the typical fillers for every word w_i.

$$EX_{role}(\langle \mathbf{w_1}, \mathbf{w_2} \rangle) = f(EX_{role_1}(\mathbf{w_1}), EX_{role_2}(\mathbf{w_2})). \tag{4}$$

Once the expectation vector has been calculated, the $filler$ fit for a $role$ can be computed by measuring the cosine similarity between the $filler$ and the expectations vector. For example, the procedure to estimate how likely is *burglar* as a patient of *the policeman arrested the...* is the following:

1. we first build a prototype vector out of the embeddings of nouns typically co-occurring with the agent *policeman*;
2. then we build another prototype vector out of the embeddings of typical patients of the verb *arrest*;
3. we combine the prototype vectors with $f(x)$;

Footnote 2 continued
also due to inter-arguments typicality (e.g. the facilitation for sentences with typical agent–patient combinations in Bicknell et al. 2010).

4. at this point, we can estimate the $filler$ thematic fit by calculating its cosine similarity ($cosSim$) to the updated prototype vector:

$$EX_{PA}(\textbf{burglar}|\langle \textbf{police}, \textbf{arrest}\rangle) = cosSim(\textbf{burglar}, f(EX_{CO}(\textbf{policeman}), EX_{PA}(\textbf{arrest}))). \quad (5)$$

In Chersoni et al. (2017b), the best performing function f turned out to be the simple vector sum between prototype vectors, and thus we used vector sum for the experiments presented in Sect. 3. According to this second model, semantic coherence is conceived as the coherence between the dynamically-updated expectations for the participants of an event described by a sentence, and the fillers saturating the participant roles. In this case, the global semantic coherence, depends on how well the last sentence argument matches the expectations generated from the sentence context.

$$\theta_{e_k} = EX_{lastRole}. \quad (6)$$

In our experiments, we refer to this model as **ThetaProtoSum**.

2.2.2 The cost of unification: event salience

In our perspective, event representations are not necessarily built on the fly: Events already stored in the GEK are activated during processing and they can progressively change their activation levels, as new words are processed. Ideally, events that satisfy all the constraints imposed by the incoming words should increase their activation, becoming the "best candidates" of a retrieval operation.

In order to account for the role of event memorization and retrieval, a second score, σ_{e_k}, is used to weight 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. The activation of an event e in GEK is computed by summing the activation scores of the single lexical items cuing it (Eq. 8), which are in turn estimated with conditional probabilities of the event given each lexical item in the input (Eq. 7):

$$\sigma_i = P(e|i) = \frac{P(e,i)}{P(i)}, \quad (7)$$

$$F(\sigma_i, \sigma_j) = \sigma_{e_k} = \sigma_i + \sigma_j. \quad (8)$$

The score σ_{e_k} measures the degree to which a unified event is activated by the linguistic expressions composing it. Consequently, events that are cued by many constructions in the sentence incrementally increase their salience.

It should be pointed out that the activation mechanism not only works for fully-specified events, but also for schematic ones (i.e., a noun *student* is supposed to activate also generic *student reading* events in the GEK). When we compute the global activation scores for a sentence s_{e_k}, we sum the scores of (i) the entire event e_k, if such an event is stored in GEK; (ii) the sub-events corresponding to all the

partial combinations of the verb and its arguments. The global activation score for the sentence s_{e_k} is computed as follows:

$$\sigma e_k = \sum_{e_i \in E} \sigma_{e_i}, \tag{9}$$

where the set of events E includes both the full event e_k and all the sub-events e_i activated by the lexical items in the input sentence.

To sum up, we weigh unified events along two dimensions: internal semantic coherence (θ), and degree of activation by linguistic expressions (σ). The latter is used to estimate the importance of "ready-to-use" event structures stored in GEK and retrieved during sentence processing. Salience scores can also be used to identify missing pieces of information, such as implicit arguments. For instance, suppose that we have the sentence *The student reads the book*, with the LOCATION role left unexpressed. If *library*-related events are simultaneously cued by *student*, *read* and *book*, their score will get higher during the integration, with the result that *library* will become a highly salient (i.e., highly probable) location for the event described in the sentence. This is a piece of unexpressed information that will be recovered during sentence comprehension. on the other hand, the θ score allows us to weigh events that are not available in the Memory component. In fact, the Unification component can construct new events never observed before, thereby accounting for the ability to comprehend novel sentences representing atypical and yet possible events.

Given an input sentence s, its interpretation INT(s) is the event e_k with the highest **semantic composition weight (SCW)**, defined as follows:

$$\text{INT}(s_k) = \underset{e_k}{\text{argmax}}(\text{SCW}(e_k)), \tag{10}$$

$$\text{SCW}(e_k) = \theta_{e_k} + \sigma_{e_k}. \tag{11}$$

Finally, we model the **argument complexity** ($ArgComp$) of a sentence s_{e_k} as inversely related to the SCW of the event representing its interpretation:

$$\text{ArgComp}(s) = \frac{1}{\text{SCW}(\text{INT}(s))}. \tag{12}$$

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 higher is the sentence argument complexity. Therefore, the joint effect of the σ and θ scores is meant to capture the "balance between storage and computation" driving sentence processing (Baggio and Hagoort 2011), and they can be considered as facilitating factors in the process of building semantic representations for the events described in natural language.

3 Case studies

We test our distributional model of argument complexity to account for the different processing costs of (i) typical vs. atypical verb-argument combinations (Sect. 3.2), and (ii) of logical metonymic vs. non-coercion sentences (Sect. 3.3).[3]

3.1 Experimental settings

First of all, we populated the DEG modelling GEK with events extracted from parsed corpora. We followed the procedure proposed in Chersoni et al. (2016b) to extract syntactic joint contexts from a concatenation of four different corpora: the Reuters Corpus Vol. 1 (Lewis et al. 2004); the ukWac and the Wackypedia Corpus (Baroni et al. 2009) and the British National Corpus (Leech 1992).[4] For each sentence, we generated a surface event representation by extracting the verb and its direct dependencies. In the present case, the dependency relations of interest are subject (NSUBJ), direct (DOBJ) and indirect object (IOBJ), infinitive and gerund complements (XCOMP), and a generic prepositional complement relation (PREPCOMP), on which we mapped all the complements introduced by a preposition. As in Chersoni et al. (2016b), we discarded all the adjectival/adverbial modifiers and just kept their heads. For instance, from the joint context *director-n-nsubj__write-v-head__article-n-dobj* we generated the event [$_E$ NSUBJ:**student** HEAD:**read** DOBJ:**book**]. For each joint context, we also generated schematic events from its dependency subsets. We extracted a total of 1,043,766 events, each including at least one of the words of the evaluation datasets.

All the lexemes in the events are represented as distributional vectors. We built a syntax-based distributional semantic model by using as targets the 20K most frequent nouns and verbs in our concatenated corpus, plus any other word occurring in the events in GEK. Words with frequency below 100 were excluded. The total number of targets is 20,560. As vector dimensions, we used the same target words, while the dependency relations are those used to build the joint contexts (e.g., the nouns NSUBJ:*chef* and DOBJ:*pizza* are examples of contexts for the verb *to cook*). Syntactic co-occurrences were weighted with Local Mutual Information (Evert 2004):

$$LMI(t, r, f) = log\left(\frac{O_{trf}}{E_{trf}}\right) \cdot O_{trf} \tag{13}$$

O_{trf} is the co-occurrence frequency of the target t, the syntactic relation r and the filler f, and E_{trf} is the expected co-occurrence frequency under independence. LMI values have been used then to rank the typical fillers for the roles in the computation of the θ components. Since our datasets are composed of agent–verb–patient triplets, we used the following approximations for semantic roles (Baroni and Lenci 2010; Lenci 2011): (i) the NSUBJ relation for the agent role; (ii) the DOBJ relation for the patient

[3] Although the architecture presented here is similar to the proposals in Chersoni et al. (2016a) and Chersoni et al. (2017a), several details of the framework have been changed, and thus the described results are different.

[4] Corpora were preprocessed with the pos-tagger described in Dell'Orletta (2009) and the dependency parser by Attardi et al. (2009).

role; (iii) a generic VERB relation for co-participants. Concretely, this relation links noun pairs that appear as subject and direct object of the same verb.

3.2 Case Study 1: modeling verb argument expectations

As a first test for our framework, we measure the argument complexity of the sentences in the **Bicknell dataset** (2010). The Bicknell dataset was collected 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 2 agent nouns that altered the scenario evoked by the agent–verb combination.

Plausible patients for each agent–verb pair were obtained with production norms, in order to generate triplets where the patient was congruent with the agent and with the verb. For each congruent triple, an incongruent one was generated by combining each verb–congruent patient pair with the other agent noun, in order to have items describing atypical situations. The final dataset includes 100 pairs of agent–verb–patient triplets, that were used to build the stimuli for a self-paced reading and an ERP experiment. For instance, subjects were presented with sentence pairs such as:

(7) 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 patient, differing for the agent. Given the agent, the patient is a preferred argument of the verb in the congruent condition, while it is way less plausible 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 knowledge about events and their typical participants. Later, Lenci (2011) was the first to use the Bicknell dataset to evaluate a distributional model for composing argument expectations on the task of assigning a higher thematic fit score to the congruent combinations than to the incongruent ones.

We interpret Bicknell's experimental data as suggesting that congruent sentences have less argument complexity than incongruent sentences. Consistently, we predict that our models will assign a higher argument complexity score to incongruent triplets than to congruent ones. Given a congruent–incongruent triple pair, we score a hit each time a model assigns a higher $ArgComp$ score to the incongruent one. Models are primarily evaluated in terms of their accuracy in this binary classification task.

3.2.1 Complexity models

For each test triple, we computed the σ and a θ scores:

– θ represents the semantic coherence of the event represented by the sentence, and is obtained by measuring the mutual typicality of its components. As we illustrated

Table 1 Examples of schematic events retrieved from the DEG to compute the σ of a given joint context

Syntactic joint context	Schematic events
NSUBJ:**general** HEAD:**assemble** DOBJ:**troop**	(NSUBJ:**general** HEAD:**assemble**)
	(HEAD:**assemble** DOBJ:**troop**)
	(NSUBJ:**general** DOBJ:**troop**)
NSUBJ:**journalist** HEAD:**write** DOBJ:**article**	(NSUBJ:**journalist** HEAD:**write**)
	(HEAD:**write** DOBJ:**article**)
	(NSUBJ:**journalist** DOBJ:**article**)

in Sect. 2.2.1, we tested two models that differ for the way they estimate semantic coherence:

1. In the $ThetaProd$ model, we computed the θ values as the product of partial thematic fit scores. Following Eq. 2, we computed θ_e for each triple as the product of (i) the thematic fit of NSUBJ given the verb HEAD, $\theta_{S,V}$; (ii) the thematic fit of DOBJ given the verb HEAD, $\theta_{O,V}$; and (iii) the thematic fit of DOBJ given NSUBJ, $\theta_{S,O}$. In particular, $\theta_{S,V}$ is the cosine between the vector of NSUBJ and the prototype vector built out of the k most salient subjects of the verb HEAD (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 DOBJ and the centroid vector built out of the k most salient direct objects of the verb HEAD (e.g., the cosine between the vector of *article* and the prototype vector of the most salient objects of *check*); and $\theta_{S,O}$ is the cosine between the vector of DOBJ and the centroid vector built out of the k most salient direct objects occurring in events where the subject is NSUBJ (e.g., the cosine between the vector of *article* and the prototype vector of the most salient objects of events whose subject is *journalist*);
2. In the $ThetaProtoSum$ model, the θ_e of each triple was computed as the similarity score between the vector of the DOBJ and the vector of the expectations for the DOBJ given NSUBJ and the verb HEAD, as in Eq. 4. Vector sum is the function that we used to combine partial prototypes in the global expectation vector for the patient (Chersoni et al. 2017b).

We identified the typical fillers for each role as the set of filler nouns with the strongest LMI score with the target word t and the relation r. Following Baroni and Lenci (2010), we set the parameter k (i.e., the number of typical fillers used to build the prototypes) to 20,

– to compute the σ score, given an event e_k, we looked for a matching syntactic joint context in our DEG repository and for schematic events matching the subchunks of e_k (some examples are shown in Table 1). For each of these events e_i, we computed the activation score by using Eqs. 7 and 8. Partial scores were then summed with Eq. 9 to obtain the global σ_e.

Finally, after computing θ_e and σ_e for each of our test triplets, we used Eqs. 10, 11, and 12 to derive the final $ArgComp$ scores.

3.2.2 Baseline models

Besides our models of argument complexity, we computed two baselines inspired by the early models of compositional distributional semantics. Mitchell and Lapata (2010) proposed two simple models for vector composition. Given the vectors of the word u and the word v, the vector representation of the expression p that they compose is computed as follows:

- in the simplified additive model (Sum):

$$\mathbf{p} = \alpha\mathbf{u} + \beta\mathbf{v}, \tag{14}$$

  where both the α and β weights are set to 1 (i.e, the output vector is the component-wise sum of the input ones);
- in the pointwise multiplicative model ($Product$):

$$p_i = u_i \cdot v_i. \tag{15}$$

Despite their simplicity, such models turned out to be extremely efficient and competitive in a wide variety of compositionality-related tasks (Rimell et al. 2016). For each triple in our dataset, we used Sum and $Product$ to build a vector representation of the patient expectations given the agent–verb combination of each dataset triple. Then, we measured the cosine similarity between the output vector and the patient one, scoring a hit whenever the score was higher for the congruent condition than for the incongruent one. The principle is the same of the *ThetaProtoSum* model: The fit of the expectations is assessed in terms of similarity between the vector of the last argument to be predicted and a vector representing the previous context, the difference being that the baseline models do not have information about typical role fillers and simply combine the vectors of the verb and its agent.

Another baseline model is based on the notion of *Surprisal*. After extracting all the subject–verb–object triples, we computed the probabilities of the trigrams and of the subject–verb bigrams with Add-One Smoothing (Jurafsky and Martin 2014). For each triple t, *Surprisal* estimates were then computed as follows:

$$Surprisal(t) = -\log_2 P(nsubj_t, verb_t, dobj_t | nsubj_t, verb_t), \tag{16}$$

where $nsubj_t$, $verb_t$ and $dobj_t$ are, respectively, the agent, the verb and the patient of t. The model accuracy is computed as the percentage of atypical triples to which it assigns a higher surprisal score.

Our models are also compared with the best configuration in Lenci (2011), that is the Product model (PROD- L11). Such model is based on the Distributional Memory data and estimates thematic fit by composing a prototype for the expectations on the patient, given the agent and the verb. In PROD- L11, a single prototype for the patient slot is built by updating the typicality scores: If a filler f has a score α_{subj} given the agent and a score α_{verb} given the verb, its typicality will be computed as $\alpha_{subj} * \alpha_{verb}$ and the prototype is built out of the 20 top fillers in the updated ranking. This way, arguments that are not compatible with both the verb and the agent are filtered out.

Table 2 Model accuracy and coverage for the classification task on the Bicknell dataset

Model	Hits/accuracy	Coverage
Random	50%	100/100
Sum	62%	100/100
Product	81.25%	96/100
ThetaProd	76.2%	84/100
ThetaProtoSum	70%	100/100
Surprisal	65%	65/100
PROD- L11	73.8%	84/100

3.2.3 Results on the Bicknell dataset

All models except for the *Sum* baseline differentiate between the two conditions. The Wilcoxon rank sum test on the output scores of the different models reveals that:

- the *ArgComp* scores assigned by *ThetaProtoSum* to the incongruent condition are significantly higher ($p < 0.05$);
- the *ArgComp* scores assigned by *ThetaProd* to the incongruent condition are significantly higher ($p < 0.01$);
- the thematic fit scores assigned by the baseline *Product* to the incongruent condition are significantly lower ($p < 0.01$).[5]

Perhaps surprisingly, the simple *Product* baseline manages to obtain the best accuracy in the binary classification task (cf. Table 2). This confirms that it is difficult to beat baselines based on simple vector operations in many compositionality-related tasks, a finding reported also by other studies on compositional distributional models (Mitchell and Lapata 2010; Rimell et al. 2016). Moreover, it has been noticed that vector multiplication eases the problem of lexical ambiguity, since dimensions that are inconsistent with the more appropriate meaning in context are filtered out. This could explain the particularly strong performance of this baseline. Still, despite being outperformed, our models also achieve high levels of accuracy and assign significantly different scores to the two conditions.

We consider the performance of *ThetaProd* to be particularly satisfactory, as it manages to outperform the original model of expectations update by Lenci (2011), when tested on the covered triplets (73.8%).[6] Moreover, its classification accuracy does not differ significantly from the one of the best-performing *Product* baseline ($p = 0.4$), while the same baseline retains a marginally significant advantage over the other complexity model, *ThetaProtoSum* ($p < 0.1$).[7] Compared to the other baselines, its advantage over *Sum* is significant at $p < 0.05$, while the difference with the *Surprisal* baseline is only marginally significant ($p < 0.1$).

[5] The scores of the baselines are not reversed as the *ArgComp* ones and they are comparable to the thematic fit scores of the θ component. Thus, the task for the baselines is to assign *lower* scores to the incongruent condition.

[6] The accuracy score has been provided by the author himself.

[7] p-values have been computed with the $\tilde{\chi}^2$ test.

Table 3 Accuracy scores for the two complexity models without the σ component and the accuracy loss with respect to the full model

Model	Hits/accuracy	Diff Full Model
ThetaProd	70.2%	**-6%**
ThetaProtoSum	70%	0%

Concerning the coverage of our models, we should also mention that for several of the triplets in the dataset (48 out of 200) the contribution of the σ component was null, as no matching joint context was retrieved from the DEG. Moreover, a syntactic joint context for the entire event could be retrieved for only 22 out of the 200 triplets. Another important point is that the task of composing and updating argument expectations is generally addressed by means of thematic fit models (Lenci 2011; Chersoni et al. 2017b) corresponding to our θ component. Thus, one might wonder if it is worth making the model more complex by introducing the extra parameter σ.

Table 3 shows the results for our complexity models after excluding σ scores from the computation. The accuracy of the *ThetaProtoSum* model remains unchanged, meaning that the direct retrieval of events from the DEG does not contribute to the correct classification of the triplets. On the other hand, the accuracy of *ThetaProd* slightly drops, and this means that the two components, in this version of the model, do not classify correctly exactly the same triplets. Although the difference (also considering the small size of the dataset) is too small to reach significance, the contribution of the two components seems to be more balanced in *ThetaProd*. From these data, it seems clear that an implementation of the memory component based only on textual corpora suffers from data sparsity (a problem that is shared with *Surprisal* models, even when smoothed), and future developments of argument complexity models will have to take this factor into account.

3.3 Case study 2: logical metonymy

In the second case study, we test our distributional approach to argument complexity on two different tasks: (i) modeling the reading times of logical metonymic sentences, (ii) and predicting the covert event that is implicitly recovered as part of their interpretation (cf. Sect. 1). For our experiments, we used two datasets created for previous psycholinguistic studies: the **McElree dataset** (McElree et al. 2001) and the **Traxler dataset** (Traxler et al. 2002). Each dataset includes three different experimental conditions, by contrasting constructions requiring a type-shift with those requiring normal composition:

(8) a. The author was starting the book.
b. The author was writing the book.
c. The author was reading the book.

Sentence (8-a) corresponds to the metonymic condition (MET), while sentences (8-b) and (8-c) correspond to non-metonymic constructions, with the difference that (8-b) contains a typical event given the subject and the object (HIGH_TYP), whereas (8-c) expresses a plausible but less typical event (LOW_TYP). The McElree dataset

was created for the self-paced reading study by McElree et al. (2001), and includes 99 sentences arranged into 33 triplets like (8), while the Traxler dataset was used in the eye-tracking experiment by Traxler et al. (2002) and contains 108 sentences (36 triplets). Three triplets of the McElree datasets were discarded, because some of their words had very low frequency in the training corpora.

3.3.1 Modeling the processing times of logical metonymy

The models have been tested on the triplets corresponding to the agent–verb–patient combination of the original datasets and the σ and θ scores have been computed like in Case Study 1. We predict that our models $ThetaProd$ and $ThetaProtoSum$ will assign *higher $ArgComp$ scores to metonymic sentences than to non-coercion sentences*, because the former do not comply with the semantic preferences of the event-selecting verb. According to Zarcone et al. (2014), it is exactly the low thematic fit between the event-selecting verb and the entity-denoting object that triggers complement coercion and that, at the same time, causes the extra processing load.

The baselines are the same we used for Case Study 1 (cf. Sect. 3.2.2) plus the following ones:

ZETAL13 Zarcone et al. (2013) proposed to model the processing costs of the same datasets by using a simpler distributional model, in which the cost of each dataset triple was computed as

$$1 - \theta(\textbf{noun}|patient, \textbf{verb}) \tag{17}$$

Therefore, this model only considers the thematic fit θ of the patient noun, without taking into account the agent filler.

SURPRISALD17 A second surprisal model, similar to the one described in the study by Delogu et al. (2017) on logical metonymy, is based on the probabilities of the trigrams composed by the verb, a determiner and the object noun. Given a trigram t, its surprisal score is computed as follows (for simplicity, we abstract away from the determiner):

$$\begin{aligned} &SurprisalD17(t) \\ &\quad = -\log_2 P(verb_t, DET, dobj_t|verb_t, DET, dobj_t), \end{aligned} \tag{18}$$

where $verb_t$ and $dobj_t$ are, respectively, the verb and the patient of the triple t, and DET is a generic determiner of the direct object. In their eye-tracking and ERP experiments, Delogu et al. (2017) reported that surprisal can fully account for the extra processing costs of logical metonymies. In other words, the expectedness of the object noun was shown to be the main determining factor of processing difficulty, without the need of postulating coercion-specific costs.

The $ThetaProd$ model turns out to be the most faithful one to the psycholinguistic results. On the McElree dataset (cf. Table 4; Fig. 2 top), the Kruskal–Wallis rank sum test revealed a main effect of the sentence types on the $SemComp$ scores assigned by

Table 4 Results of the pairwise post hoc comparisons for the three conditions on the McElree dataset (Wilcoxon rank sum test with Bonferroni correction), scores assigned by *ThetaProd*

p-values	HIGH_TYP	LOW_TYP
LOW_TYP	0.04*	–
MET	0.00046*	0.31

Table 5 Results of the pairwise post hoc comparisons for the three conditions on the Traxler dataset (Wilcoxon rank sum test with Bonferroni correction), scores assigned by *ThetaProd*

p-values	HIGH_TYP	LOW_TYP
LOW_TYP	0.31	–
MET	9.7e−06*	0.01*

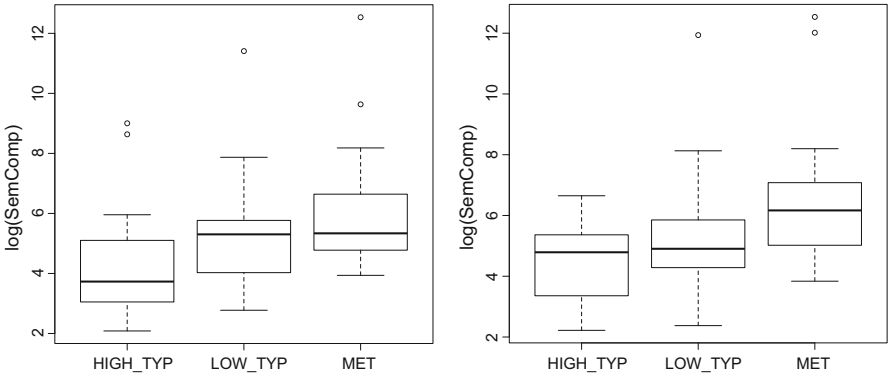

Fig. 2 *SemComp* scores for McElree (left) and Traxler (right), computed with the *ThetaProd* model

ThetaProd ($\chi^2 = 17.18$, $p < 0.001$). Post hoc tests showed that *SemComp* scores for the HIGH_TYP conditions are significantly lower than those in the LOW_TYP ($p < 0.05$) and MET conditions ($p < 0.001$). These results mirror exactly those of McElree et al. (2001) for the reading times at the type-shifted noun (both conditions engendered significantly longer reading times than the preferred condition).

A main effect of sentence types on the *SemComp* scores was also found for the Traxler dataset ($\chi^2 = 15.39$, $p < 0.001$). In their eye-tracking experiment (Experiment 1), Traxler et al. (2002) found no significant difference between HIGH_TYP and LOW_TYP conditions, but they observed higher values for second-pass and total time data in the MET condition with respect to the other two. Interestingly, the *ThetaProd* model produced similar results (cf. Table 5; Fig. 2 bottom): post hoc tests reveal no difference between non-coerced conditions, but significantly higher *SemComp* scores for metonymic sentences with respect to both the HIGH_TYP ($p < 0.001$) and the LOW_TYP condition ($p < 0.05$).

The *ThetaProtoSum* model also assigned significantly different scores to the three conditions, both in the McElree ($\chi^2 = 28.64$, $p < 0.001$) and in the Traxler dataset ($\chi^2 = 26.656$, $p < 0.001$). However, the results of this model did not reproduce so accurately the results of the experiments, as the assigned scores simply discriminate between metonymic and non-metonymic conditions in both datasets (see Tables 8, 9).

Table 6 Results of the pairwise post hoc comparisons for the three conditions on the McElree dataset (Wilcoxon rank sum test with Bonferroni correction), scores assigned by *ThetaProtoSum*

p-values	HIGH_TYP	LOW_TYP
LOW_TYP	0.195	–
MET	4.5e−07*	0.002*

Table 7 Results of the pairwise post hoc comparisons for the three conditions on the Traxler dataset (Wilcoxon rank sum test with Bonferroni correction), scores assigned by *ThetaProtoSum*

p-values	HIGH_TYP	LOW_TYP
LOW_TYP	0.68	–
MET	2.4e−07*	0.00084*

Table 8 Summary table with the results of all the pairwise comparisons on the McElree dataset for all models

Model	HIGH_TYP vs. LOW_TYP	HIGH_TYP vs. MET	LOW_TYP vs. MET
Sum	–	✓	–
Product	–	–	–
Surprisal	–	–	–
SURPRISALD17	–	✓	–
ZETAL13	✓	✓	✓
ThetaProd	✓	✓	–
ThetaProtoSum	–	✓	✓
Experiment	✓	✓	–

Checkmarks indicate significant differences, while the **Experiment** line reports the pattern found in the original experiment

This pattern is very close to the one found by ZETAL13, which discriminates between HIGH_TYP and MET ($p < 0.001$) and LOW_TYP and MET ($p < 0.01$) on both datasets. Additionally, ZETAL13 found a marginally significant difference between HIGH_TYP and LOW_TYP in the McElree dataset.

Concerning the baseline models, the original *Surprisal* (with Add-One smoothing) fails to differentiate between conditions in both datasets. *SurprisalD17*, instead, generates significantly different scores on both the McElree ($\chi^2 = 6.05$, $p < 0.05$)and the Traxler dataset ($\chi^2 = 7.02$, $p < 0.05$), but the only conditions that differ are HIGH_TYP and MET (in both cases, $p < 0.05$). Finally, both the simple DSM baselines struggle in differentiating between the three experimental conditions: for the Kruskal–Wallis test, the differences between the scores assigned by *Sum* and *Product* never reach significance, with the only exception of *Sum* on the McElree dataset ($p < 0.05$). Coming to pairwise comparisons, the pattern is different than the one reported by McElree and colleagues, since no significant difference between HIGH_TYP and LOW_TYP has been found ($p = 0.9$).

Table 9 Summary table with the results of all the pairwise comparisons on the Traxler dataset for all models

Model	HIGH_TYP vs. LOW_TYP	HIGH_TYP vs. MET	LOW_TYP vs. MET
Sum	–	–	–
Product	–	–	–
Surprisal	–	–	–
SURPRISALD17	–	✓	–
ZETAL13	–	✓	✓
ThetaProd	–	✓	✓
ThetaProtoSum	–	✓	✓
Experiment	–	✓	✓

Checkmarks indicate significant differences, while the **Experiment** line reports the pattern found in the original experiment

3.3.2 Identifying the covert event

We assume that the SR of a metonymic sentence like *The author starts the book* contains the following complex event:

(9) $[_{E_1}$ NSUBJ:**author** HEAD:**start** DOBJ:$[_{E_{cov}}$ NSUBJ:**author** HEAD:**E$_{cov}$** DOBJ:**book**]]

where E_{cov} is the covert event recovered when interpreting the sentences (e.g., *writing*). We modeled covert event retrieval as a binary classification task: *Given a set of candidate hidden events, we argue that the selected interpretation is the one that minimizes argument complexity*. This claim was tested with the following procedure:

1. for each metonymic sentence (e.g., *The author starts the book*) in the McElree and Traxler datasets, we selected as candidate covert events (E_{cov}) the verbs in the non-coercion sentences, which we refer to respectively as HIGH_TYP_EVENT (e.g., *write*) and LOW_TYP_EVENT (e.g., *read*). Therefore, we obtain quadruple pairs like the following ones:

(10) a. author start write book (HIGH_TYP_EVENT)
b. author start read book (LOW_TYP_EVENT)

2. for each sentence $SV_{met}O$, we computed SCW(e) (cf. Eq. 11) of the events composing its interpretation, that is $[_E$ S V_{met} $E_{cov}]$ and $[_{E_{cov}}$ S E_{cov} O] (i.e., we computed it for both the HIGH_TYP and the LOW_TYP quadruple in each pair);[8]
3. the model **accuracy** was computed as the percentage of test items for which SCW(E_{cov} = HIGH_TYP_EVENT) > SCW(E_{cov} = LOW_TYP_EVENT).

We compared our distributional approach with the probabilistic model introduced by Zarcone et al. (2012), and we computed the probability $P(e)$ of a candidate verb

[8] Importantly, the covert events do not contribute to the σ scores, since the corresponding verbs are not present in the linguistic input.

Table 10 Accuracy (and coverage) of the models and of the baselines on the binary classification task for covert event retrieval

Model	McElree	Traxler
Random	50% (30)	50% (36)
Sum	40% (30)	50% (36)
Product	56.6% (30)	50% (36)
ThetaProd	80% (30)	77.77% (36)
ThetaProtoSum	66% (30)	52.77% (36)
Surprisal	66.6% (30)	58.3% (36)
ZETAL12	77.7% (18)	72% (25)

Table 11 Accuracy of *ThetaProd* after the removal of σ and performance drop on the McElree and the Traxler datasets

Dataset	Accuracy	Performance drop
McElree	73.3%	**-6.7%**
Traxler	75%	**-2.7%**

as the hidden event E_{cov} as:

$$P(e) = P(verb) \cdot P(subject|verb) \cdot P(object|verb). \tag{19}$$

We refer to this model as ZETAL12. This is a generative model, since it first assumes a hidden event E_{cov} and then generates the arguments on the basis of the choice of E_{cov}. When compared with other distributional models of logical metonymy, ZETAL12 achieved the highest accuracy, but a lower coverage due to the zero-counts of many of the co-occurrences needed to compute the probabilities in (19).

The results for the covert event identification are shown in Table 10. Overall, we can observe that the *ThetaProd* model is again the best performing one, classifying correctly almost all the triplets, and it is the only one to significantly outperform a random baseline at $p < 0.05$ in both the McElree and the Traxler dataset.[9] Conversely, *ThetaProtoSum*, *Sum*, *Product* and *Surprisal* struggle in this classification task, and they barely manage to classify a few triples more than a random baseline.

The model going closer to *ThetaProd* in terms of accuracy is the reimplementation of ZETAL12. Like in the original study, this probabilistic model has very high accuracy, but it also struggles with data sparsity and has a more limited coverage. Again, we tested the *ThetaProd* model by removing the σ component, in order to assess its contribution to the classification task. Once again, the contribution of the σ component is limited to few triplets, especially on the Traxler dataset that includes several rare words (cf. Table 11). It is the θ component to play the crucial role in the covert event prediction, while for unusual and rare events, there is simply no matching joint context that can be retrieved from the DEG representing *GEK*.

As a final experiment, we wanted to test the claim by Zarcone et al. (2013, 2014), according to which thematic fit estimation is the mechanism responsible for the triggering of logical metonymy. Their hypothesis was that the recovery of the implicit event

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

could be a consequence of the dispreference of the verb for the entity-denoting argument. In our framework, this corresponds to saying that the low thematic fit between verb and patient triggers a retrieval operation with the aim of increasing the semantic coherence of the event represented in the SR. To test this claim, we compared the θ scores of the events containing the HIGH_TYP covert event (i.e., $[_E$ S V_{met} $E_{cov}]$ + $[_{E_{cov}}$ S E_{cov} O]) and the corresponding MET event (i.e., $[_E$ S V_{met} O]), predicting that the former events are more semantically coherent than the latter.[10] This hypothesis turned out to be correct: According to the Wilcoxon rank sum test, both in the McElree ($W = 199$, $p < 0.01$) and in the Traxler dataset ($W = 157$, $p < 0.01$) the θ of the structures with the covert events are significantly higher.

4 Discussion

We introduced a framework for argument complexity relying on the two components of Memory and Unification, as in the MUC framework by Hagoort (2013). The first refers to the storage of GEK that we represent by means of the corpus-derived DEG, whereas the second concerns the constraint-driven combination of the units stored in the DEG into more complex structures. Our hypothesis is that GEK stores information about typical events and participants, and that this knowledge allows speakers to anticipate the upcoming linguistic input during sentence processing. Human lexical knowledge, as argued by several modern theories of language processing (Libben 2005; Marzi and Pirrelli 2015), does not seem to be organized to minimize storage, but rather to maximize processing efficiency.

Words work as cues to GEK (Elman 2014), and the recovered information is dynamically unified to build a representation of the events that natural language sentences are likely to communicate. Differently from other approaches, mainly looking at syntactic factors, we focused on the semantics of the events described by natural language sentences and used syntax only to identify aspects of their structure. However, a complete model of processing complexity could separately represent the relevant information for each linguistic domain by means of different constraints (Blache 2016), and domain-specific complexity indexes could be somehow combined and integrated in order to account for the different complexity sources.

In the proposed DSM-based implementation, event representations are weighted along two different dimensions:

- the **semantic coherence** θ of the unified event, which depends on the mutual typicality between the participants and is computed with a distributional model of thematic fit;
- the **activation by lexical items or salience** σ, which corresponds to the activation strength of GEK events cued by lexical items. Activation values are modeled as simple conditional probability scores, and the global activation of an event is computed by taking into account also the contribution of schematic events.

[10] Since the computation of the two θs in $ThetaProd$ requires a different number n of factors, the scores have been normalized by elevating them to the power of $1/n$.

An important assumption of our model is that the argument complexity of a sentence is inversely-related to these two factors: (i) the activation strength of a corresponding event stored in GEK, and (ii) the mutual typicality of its participants, resulting in a more predictable situation. In our experiments, we compared the predictions of our model with the findings of some psycholinguistic studies. The most successful version of the model turned out to be $ThetaProd$, which computed the θ component as the product of the single event-participant thematic fit scores. We argue that this approach has several elements of strength:

- it achieved a competitive performance on the binary classification task for the update of context-sensitive argument typicality, evaluated on the data by Bicknell et al. (2010), being outperformed only by a strong $Product$ baseline, which however obtain suboptimal performances in the other tasks;
- in modeling the processing cost of logical metonymy observed in the studies by McElree et al. (2001) and Traxler et al. (2002), $ThetaProd$ closely reproduced the behavioral data showing significant differences between the three experimental conditions (typical, non-typical and metonymic event);
- in retrieving the covert event of logical metonymy, which turned out to be difficult for all the models, it achieved the best performance and was the only system managing to significantly outperform a random baseline. Moreover, it does not suffer from the coverage problems of probabilistic models (Zarcone et al. 2012) ;
- the θ component assigns significantly higher scores to metonymic verbs (e.g., *finish*) with a non-event denoting direct object (e.g., *book*) than to the corresponding structure *after* the integration of the covert event. This is coherent with the hypothesis by Zarcone et al. (2013, 2014), according to which the covert event retrieval is triggered by a low thematic fit between verb and object, and it is aimed at "repairing" the low degree of semantic coherence of the metonymic structure;
- finally, the addition of the σ component leads to some improvement (although not significant) over the thematic fit model alone (θ), making us think that the action of the two components can be somehow considered as complementary.

An actual limit of the model is the coverage of the σ component, which was found to be low on all datasets. On the one hand, this could perfectly make sense, as it is difficult to think that a semantic memory component could store all possible events. In most cases, it is likely that the semantic representations for the events have to be built from scratch. On the other hand, it would be desirable for future extensions of such a model to implement some sort of generalization on the basis of the similarity between the arguments. For example, there might be no distributional information stored for the event of a policeman arresting a *burglar*, but there might be one for a policeman arresting a *crook*. The ability to generalize, by recognizing the similarity between the two situations and adapting the stored representation to the new event, would be extremely useful for increasing the contribution of the σ component. At the same time, the results of our experiments confirm that argument complexity and its online processing effects need to be explained within a general model of the incremental and compositional construction of the semantic representations of sentences describing previously unseen events, which is the very essence of natural language productivity.

5 Conclusions and future work

In this work, we have presented a distributional model of argument complexity and we tested it on the tasks of accounting for sentence typicality and logical metonymy resolution. In our view, these are two aspects of the same phenomenon, as in both cases the argument typicality determines processing complexity.

On the computational side, one of our models showed a nice capacity of handling both phenomena in two different psycholinguistic datasets, proving to be more general than previous approaches. It should be pointed out, however, that our datasets were not the ideal ones for an exhaustive comparison between the models, given their small size and the relatively simple structure, which has been modeled as subject–verb–object triplets. As we anticipated in the introduction, we treated the problem of semantic complexity mainly in relation to the problem of argument typicality, but this entails ruling out several, potential sources of complexity, such as more complex event structures (i.e., events including also roles like instruments and locations), the presence of argument modifiers, and semantic relatedness effects due to the sentence or wider discourse context. Many current approaches to the estimation of argument typicality also limit themselves to relatively easy tasks, and one of the main reasons is the well-known scarcity of benchmark datasets (Vassallo et al. 2018). Hopefully, the joint effort of NLP and psycholinguistic research in the next years will produce more robust benchmarks, built with the goal of evaluating argument complexity models on a wider variety of structures, and taking into account semantic complexity stemming from different linguistic domains (Blache 2011, 2016).

References

Attardi, G., Dell'Orletta, F., Simi, M., & Turian, J. (2009). Accurate dependency parsing with a stacked multilayer perceptron. In *Proceedings of EVALITA*.

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes, 26*(9), 1338–1367.

Baggio, G., Van Lambalgen, M., & Hagoort, P. (2012). The processing consequences of compositionality. In *The Oxford handbook of compositionality* (pp. 657–674). Oxford: Oxford University Press.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics, 36*(4), 673–721.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation, 43*(3), 209–226.

Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language, 63*(4), 489–505.

Blache, P. (2011). Evaluating language complexity in context: New parameters for a constraint-based model. In *CSLP-11, workshop on constraint solving and language processing*.

Blache, P. (2016). Representing syntax by means of properties: A formal framework for descriptive approaches. *Journal of Language Modelling, 4*(2), 183–224.

Chersoni, E., Blache, P., & Lenci, A. (2016a). Towards a distributional model of semantic complexity. In *Proceedings of the COLING workshop on computational linguistics for linguistic complexity (CL4LC)*.

Chersoni, E., Santus, E., Lenci, A., Blache, P., & Huang, C. R. (2016b). Representing verbs with rich contexts: An evaluation on verb similarity. In *Proceedings of EMNLP*.

Chersoni, E., Lenci, A., & Blache, P. (2017a). Logical metonymy in a distributional model of sentence comprehension. In *Proceedings of *SEM*.

Chersoni, E., Santus, E., Blache, P., & Lenci, A. (2017b). Is structure necessary for modeling argument expectations in distributional semantics? In *Proceedings of IWCS*.
Chersoni, E., Santus, E., Pannitto, L., Lenci, A., Blache, P., & Huang, C. R. (2019). A structured distributional model of sentence meaning and processing. *Natural Language Engineering, 25*(4), 483–502.
Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research, 12*, 2493–2537.
Dell'Orletta, F. (2009). Ensemble system for part-of-speech tagging. In *Proceedings of EVALITA*.
Delogu, F., Crocker, M. W., & Drenhaus, H. (2017). Teasing apart coercion and surprisal: Evidence from eye-movements and ERPs. *Cognition, 161*, 46–59.
Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science, 33*(4), 547–582.
Elman, J. L. (2014). Systematicity in the lexicon: On having your cake and eating it too. In P. Calvo & J. Symons (Eds.), *The architecture of cognition: Rethinking Fodor and Pylyshyn's systematicity challenge*. Cambridge, MA: The MIT Press.
Erk, K. (2007). A simple, similarity-based model for selectional preferences. In *Proceedings of ACL*.
Erk, K., Padó, S., & Padó, U. (2010). A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics, 36*(4), 723–763.
Evert, S. (2004). The statistics of word cooccurrences: Word pairs and collocations. PhD Thesis, University of Stuttgart.
Frisson, S., & McElree, B. (2008). Complement coercion is not modulated by competition: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(1), 1–11.
Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, Language, Brain* (pp. 95–26). Cambridge, MA: MIT Press.
Greenberg, C., Sayeed, A. B., & Demberg, V. (2015). Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering. In *Proceedings of HLT-NAACL*.
Hagoort, P. (2013). MUC (memory, unification, control) and beyond. *Frontiers in Psychology, 4*, 416.
Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL-HLT*.
Hale, J. (2016). information-theoretical complexity metrics. *Language and Linguistics Compass, 10*(9), 397–412.
Hare, M., Jones, M., Thomson, C., Kelly, S., & McRae, K. (2009). Activating event knowledge. *Cognition, 111*(2), 151–167.
Heinzerling, B., Moosavi, N. S., & Strube, M. (2017). Revisiting selectional preferences for coreference resolution. In *Proceedings of EMNLP*.
Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing* (Vol. 3). London: Pearson.
Kamp, H. (2013). *Meaning and the dynamics of interpretation: Selected papers by Hans Kamp*. Leiden: Brill.
Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience, 31*(5), 602–616.
Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*(4427), 203–205.
Lapata, M., & Lascarides, A. (2003). A probabilistic account of logical metonymy. *Computational Linguistics, 29*(2), 261–315.
Lascarides, A., & Copestake, A. (1998). Pragmatics and word meaning. *Journal of Linguistics, 34*, 378–414.
Lauwers, P., & Willems, D. (2011). Coercion: Definition and challenges, current approaches, and new trends. *Linguistics, 49*(6), 1219–1235.
Lebani, G. E., & Lenci, A. (2018). A distributional model of verb-specific semantic roles inferences. In *Language, cognition, and computational models* (pp. 118–158). Cambridge: Cambridge University Press.
Leech, G. N. (1992). 100 Million words of English: The British National Corpus (BNC). In *Second language research*.
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*.
Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics, 4*, 151–171.
Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research, 5*(Apr), 361–397.
Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29*(3), 375–419.

Libben, G. (2005). Everything is psycholinguistics: Material and methodological considerations in the study of compound processing. *Canadian Journal of Linguistics/Revue Canadienne de Linguistique, 50*(1–4), 267–283.
Marzi, C., & Pirrelli, V. (2015). A neuro-computational approach to understanding the mental lexicon. *Journal of Cognitive Science, 16*(4), 493–534.
Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., & McRae, K. (2011). Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(4), 913.
McElree, B., Traxler, M. J., Pickering, M. J., Seely, R. E., & Jackendoff, R. (2001). Reading time evidence for enriched composition. *Cognition, 78*, B17–B25.
McRae, K., & Matsuki, K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass, 3*(6), 1417–1429.
McRae, K., Spivey-Knowlton, M., & Tanenhaus, M. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 38*(3), 283–312.
McRae, K., Hare, M., Elman, J. L., & Ferretti, T. (2005). A basis for generating expectancies for verbs from nouns. *Memory and Cognition, 33*(7), 1174–1184.
Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science, 34*(8), 1388–1429.
Pustejovsky, J., & Batiukova, O. (2019). *The lexicon*. Cambridge: Cambridge University Press.
Resnik, P. (1997). Selectional preference and sense disambiguation. In *Tagging text with lexical semantics: Why, what, and how?*
Rimell, L., Maillard, J., Polajnar, T., & Clark, S. (2016). RELPRON: A relative clause evaluation dataset for compositional distributional semantics. *Computational Linguistics, 42*(4), 661–701.
Roth, M., & Lapata, M. (2015). Context-aware frame-semantic role labeling. *Transactions of the Association for Computational Linguistics, 3*, 449–460.
Santus, E., Chersoni, E., Lenci, A., & Blache, P. (2017). Measuring thematic fit with distributional feature overlap. In *Proceedings of EMNLP*.
Sayeed, A., Demberg, V., & Shkadzko, P. (2015). An exploration of semantic features in an unsupervised thematic fit evaluation framework. *Italian Journal of Computational Linguistics, 1*(1), 31–46.
Tilk, O., Demberg, V., Sayeed, A., Klakow, D., & Thater, S. (2016). Event participant modelling with neural networks. In *Proceedings of the EMNLP*.
Traxler, M. J., Pickering, M. J., & McElree, B. (2002). Coercion in sentence processing: Evidence from eye-movements and self-paced reading. *Journal of Memory and Language, 47*(4), 530–547.
Vassallo, P., Chersoni, E., Santus, E., Lenci, A., & Blache, P. (2018). Event knowledge in sentence processing: A new dataset for the evaluation of argument typicality. In *Proceedings of the LREC workshop on linguistic and neuro-cognitive resources (LiNCR)*.
Zapirain, B., Agirre, E., Marquez, L., & Surdeanu, M. (2013). Selectional preferences for semantic role classification. *Computational Linguistics, 39*(3), 631–663.
Zarcone, A., Utt, J., & Padó, S. (2012). Modeling covert event retrieval in logical metonymy: Probabilistic and distributional accounts. In *Proceedings of the NAACL workshop on cognitive modeling and computational linguistics*.
Zarcone, A., Lenci, A., Padó, S., & Utt, J. (2013). Fitting, not clashing! A distributional semantic model of logical metonymy. In *Proceedings of IWCS*.
Zarcone, A., Padó, S., & Lenci, A. (2014). Logical metonymy resolution in a words-as-cues framework: Evidence from self-paced reading and probe recognition. *Cognitive Science, 38*(5), 973–996.
Zhang, H., Ding, H., & Song, Y. (2019). SP-10K: A large-scale evaluation set for selectional preference acquisition. In *Proceedings of ACL*.
Zhang, H., Bai, J., Song, Y., Xu, K., Yu, C., Song, Y., Ng, W., & Yu, D. (2020.) Multiplex word embeddings for selectional preference acquisition. arXiv preprint arXiv:200102836.