Logical Metonymy in a Distributional Model of Sentence Comprehension

Emmanuele Chersoni
Aix-Marseille University
emmanuelechersoni@gmail.com

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

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

Abstract

In theoretical linguistics, *logical metonymy* is defined as the combination of an event-subcategorizing verb with an entity-denoting direct object (e.g., *The author began the book*), so that the interpretation of the VP requires the retrieval of a covert event (e.g., *writing*). Psycholinguistic studies have revealed extra processing costs for logical metonymy, a phenomenon generally explained with the introduction of new semantic structure. In this paper, we present a general distributional model for sentence comprehension inspired by the Memory, Unification and Control model by Hagoort (2013, 2016). We show that our distributional framework can account for the extra processing costs of logical metonymy and can identify the covert event in a classification task.

1 Logical Metonymy: Psycholinguistic Evidence and Computational Modeling

The interpretation of so-called *logical metonymy* (e.g, *The student begins the book*) has received an extensive attention in both psycholinguistic and linguistic research. The phenomenon is extremely problematic for traditional theories of compositionality (Asher, 2015) and is generally explained as a type clash between an event-selecting metonymic verb (e.g., *begin*) and an entity-denoting nominal object (e.g., *the book*), which triggers the recovery of a hidden event (e.g., *reading*). Past research work brought extensive evidence that such metonymic constructions also determine extra processing costs during online sentence comprehension (McElree et al., 2001; Traxler et al., 2002), although such evidence is not uncontroversial (Falkum, 2011). According to Frisson and McElree (2008), event recovery is triggered by the type clash, and the extra processing load is due to ”the deployment of operations to construct a semantic representation of the event”. Thus, logical metonymy raises two major questions: i.) How is the hidden event recovered? ii.) What is the relationship between such mechanism and the increase in processing difficulty?

One of the first accounts of the phenomenon dates back to the works of Pustejovsky (1995) and Jackendoff (1997), which assume that the covert event is retrieved from complex lexical entries consisting of rich knowledge structures (Pustejovsky's *qualia roles*). For example, the representation of a noun like *book* includes telic properties (the purpose of the entity, e.g. *read*) and agentive properties (the mode of creation of the entity, e.g. *write*). The predicate-argument type mismatch triggers the retrieval of a covert event from the object noun qualia roles, thereby producing a semantic representation equivalent to *begin to write the paper* (see also the discussion in Traxler et al. (2002)).

On the one hand, the lexicalist explanation is very appealing, since it accounts for the existence of default interpretations of logical metonymies (e.g. *begin the book* is typically interpreted as *begin reading/writing the book*). On the other hand, Lascarides and Copestake (1998) and more recently Zarcone et al. (2014) show that qualia roles are simply not flexible enough to account for the wide variety of interpretations that can be retrieved. These are in fact affected by the subject choice, the general syntactic and discourse context, and by our world knowledge. [1]

[1] Consider the classical example from Lascarides and Copestake (1998): *My goat eats anything. He really enjoys*

*Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 168–177,
Vancouver, Canada, August 3-4, 2017. ©2017 Association for Computational Linguistics

An alternative view on logical metonymy has been proposed in the field of relevance-theoretic pragmatics (Sperber and Wilson, 1986; Carston, 2002). According to studies such as de Almeida (2004), de Almeida and Dwivedi (2008) and Falkum (2011), the metonymy resolution process is driven by post-lexical pragmatic inferences, relying on both general world knowledge and discourse context. The 'pragmatic hypothesis' allows for the necessary flexibility in the interpretation of logical metonymies, since the range of the potential covert events is not constrained by the lexical entry, but only by the hearer's expectations of the optimal relevance of the utterance. However, as pointed out by Zarcone and Padó (2011), the pragmatic account is not precise with respect to the mechanism and to the type of knowledge involved in the process of metonymy resolution. Moreover, it tends to disregard the fact that there are default interpretations that are activated in neutral, less informative contexts.

More recently, Zarcone and Padó (2011) and Zarcone et al. (2014) brought experimental evidence for the role of Generalized Event Knowledge (GEK) (McRae and Matsuki, 2009) in the interpretation of logical metonymies. The authors refer to a long trend of psycholinguistic studies (McRae et al., 1998; Altmann, 1999; Kamide et al., 2003; McRae et al., 2005; Hare et al., 2009; Bicknell et al., 2010), which show that speakers quickly make use of their rich event knowledge during online sentence processing to build expectations about the upcoming input.[2] The experiments on German by Zarcone et al. (2014) show that the subjects combine the linguistic cues in the input to activate typical events the sentences could refer to. Given an agent-patient pair, if the covert event is typical for that specific argument combination, it is read faster and it is more difficult to inhibit in a probe recognition task. The authors explained their results in the light of the words-as-cues paradigm (Elman, 2009, 2014), which claims that the words in the mental lexicon are cues to event knowledge modulating language comprehension in an incremental fashion.

Research in computational semantics has focused on two different aspects of the phenomenon: the first one is the retrieval of the covert event, which has been approached by means of either probabilistic methods (Lapata and Lascarides, 2003; Lapata et al., 2003; Shutova, 2009) or of distributional similarity-based thematic fit estimations (Zarcone et al., 2012), whereas the second aspect concerns modeling the experimental data about processing costs. Zarcone et al. (2013) showed that a distributional model of verb-object thematic fit can reproduce the reading times differences in the experimental conditions found by McElree et al. (2001) and Traxler et al. (2002). Their merits notwithstanding, a limit of the former studies is that they did not try to build a single model to account for both aspects involved in logical metonymy.

The goal of this paper is twofold. First of all, we present a **general distributional model of sentence comprehension** inspired by recent proposals in neurocognitive sciences (Section 2). Secondly, we introduce a **semantic composition weight** that is used to model the reading times of metonymic sentences reported in previous experimental studies and to predict the covert event in a binary classification task (Section 3).

2 A Distributional Model of Sentence Comprehension

The model we present includes a **Memory component**, containing distributional information activated by lexical items, and a **Unification component**, which combines the items in Memory to form a coherent semantic representation of the sentence.[3] This architecture is directly inspired by Memory, Unification and Control (MUC), proposed by Peter Hagoort as a general model for the neurobiology of language (Hagoort, 2013, 2016). MUC incorporates three main functional components: i.) *Memory* corresponds to linguistic knowledge stored in long-term memory; ii.) *Unification* refers to the assembly in working memory of the constructions stored in Memory into larger structures, with contributions from the context; iii.) *Control* is responsible for relating language to joint action and social interaction. Similarly to

your book (= eating). The event retrieval cannot be explained in terms of qualia structures, as it is unlikely that the lexical entry for *book* includes something related to *eating*-events.

[2] It should be pointed out that, unlike relevance theory which conceives world knowledge and linguistic knowledge as separate modules, GEK includes both linguistic and extralinguistic information.

[3] A previous version of this model has already been introduced in Chersoni et al. (2016a), the main difference being the way the complexity score component based on Memory was computed (see section 5 and 6 of the 2016 paper). Moreover, the model was applied to a different task (i.e., the computation of context-sensitive argument typicality).

MUC, we argue that the comprehension of a sentence is an incremental process driven by the goal of constructing a coherent semantic representation of the event the speaker intends to communicate. Our model rests on the following assumptions:

- the Memory component contains information about events and their typical participants, which is derived from both first-hand experience and linguistic experience. Following McRae and Matsuki (2009), we call this information **Generalized Event Knowledge** (GEK). In this paper we restrict ourselves to the 'linguistic' subset of GEK (henceforth GEK_L), which we model with distributional information extracted from corpora;
- during sentence processing, lexical items activate portions of GEK_L, and the Unification component composes them into a coherent representation of the event expressed by the sentence;
- the event representation is assigned a **semantic composition weight** on the basis of i) the availability and salience of information stored in GEK_L and activated by the linguistic input; ii) the semantic coherence of the unified event, depending in turn on the mutual typicality of the event participants;
- a sentence interpretation is the event with the highest semantic composition weight, that is the event that best satisfies the semantic constraints coming from lexical items and the contextual information stored in GEK_L.

Sentence comprehension therefore results from a "balance between storage and computation" (Baggio and Hagoort, 2011; Baggio et al., 2012) that simultaneously accounts for the unlimited possibility to understand new sentences, which are constructed by means of Unification, and for the processing advantage guaranteed by the retrieval from Memory of "ready-to-use" information about typical events and situations.

Crucially, we argue that logical metonymy interpretation shares this same mechanism of on-line sentence processing and that the covert event is i.) an event retrieved from GEK_L that is strongly activated by the lexical items, ii.) and with a high degree of mutual semantic congruence with the other arguments in the sentence. Therefore, there is no formal difference between simple and enriched forms of compositionality (Jackendoff, 1997), both being instances of the same general model of sentence processing.

2.1 The Memory Component: A Distributional Model of GEK_L

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_L, and σ_i is an activation score computed as the conditional probability $P(e|w_i)$, which quantifies the 'strength' with which the event is activated by w_i.

We represent **events** in GEK_L as feature structures specifying participants and roles, and we extract this information from parsed sentences in corpora: the attributes are syntactic dependencies, which we use as a surface approximation of semantic roles, and the values are distributional vectors of dependent lexemes.[4] For example, from the sentence *The student reads a book* we extract the following event representation:

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

Events in GEK_L 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_L to contain schematic (i.e., underspecified) events too. For instance, from the sentence *The student reads a book* we also generate **schematic events** such as $[_{EVENT}\ \text{NSUBJ}{:}\overrightarrow{student}\ \text{DOBJ}{:}\overrightarrow{book}]$, obtained by abstracting over one or more of the instantiated attribute values. Such representation describes an underspecified event schema involving a student and a book, which can be instantiated by different activities (e.g., *reading*, *borrowing*, etc.). According to this view, GEK_L is not a flat list of events, but a structured repository of prototypical knowledge about event contingencies.

It is worth remarking that the events in GEK_L are complex symbolic structures including distributional representations of the event head and its participants. Events in GEK_L are therefore modeled like a sort of **semantic frames** whose elements are distributional vectors.[5]

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

[5] Unlike traditional semantic frames, our events are satu-

2.2 The Unification Component: Building Semantic Representations

Language can be seen as a set of instructions that the comprehender uses to create a representation of the situation that is being described by the speaker. In our framework, we make use of **situation models** (henceforth SMs),[6] defined as data structures that contain a representation of the event currently being processed (Zwaan and Radvansky, 1998). Comprehension always occurs within the context of an existing SM: during online sentence processing, lexical items cue portions of GEK_L and the SM is dynamically updated by unifying its current content with the new information. In this perspective, the goal of sentence comprehension consists in recovering (reconstructing) the event e that the sentence is most likely to describe (Kuperberg, 2016). The event e is the event that best satisfies all the constraints set by the lexical items in the sentence and by the active SM.[7]

Let $w_1, w_2, \ldots, w_n$ be an input linguistic sequence (e.g., a sentence or a discourse) that is currently being processed. 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}:

1. the GEK_L associated with w_{i+1} in the lexicon, $GEK_L[w_{i+1}]$, is activated;

2. $GEK_L[w_{i+1}]$ 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_L cued by the linguistic elements that are composed:

$$F(SM_i, GEK_L[w_{i+1}]) = SM_{i+1} \qquad (1)$$

The composition function is responsible for two distinct processes:

- F **unifies** two event feature structures into a new event, provided that the attribute-value features of the input events are compatible. Here is an example of unification:

$$[_{EVENT}\ \text{NSUBJ}:\overrightarrow{mechanic}\ \ \text{DOBJ}:\overrightarrow{engine}] \sqcup [_{EVENT}\ \text{NSUBJ}:\overrightarrow{mechanic}\ \text{HEAD}:\overrightarrow{check}] = [_{EVENT}\ \text{NSUBJ}:\overrightarrow{mechanic}\ \text{HEAD}:\overrightarrow{check}\ \text{DOBJ}:\overrightarrow{engine}]$$

  The event of a *mechanic* performing an action on an *engine* and the event of a *mechanic checking* something are unified into a new 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 and its salience given the lexical cues activating it.

The score θ_{e_k} quantifies 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 the following sentences:

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

The event represented in (1-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 (1-b) have a low level of mutual typicality, thereby resulting into an event with much lower semantic coherence. Although the sentence is perfectly understandable, it sounds a little weird because it depicts an unusual situation.

We measure the mutual typicality of the components by extending the notion of **thematic fit**, which is normally used to measure the congruence of a predicate with an argument (McRae et al., 1998). In our case, instead, thematic fit is a general measure of the semantic typicality or congruence among event participants. Extending the approach by Erk et al. (2010), thematic fit is measured with vector cosine in the following way:

> $\theta(\overrightarrow{a}|s_i, \overrightarrow{b})$ (the thematic fit of $\overrightarrow{a}$ given $\overrightarrow{b}$ and the role s_i) is the cosine between $\overrightarrow{a}$ and the prototype vector built out of the k top values $\overrightarrow{c_1}, \ldots, \overrightarrow{c_k}$, such that $s_i{:}\overrightarrow{c_z}$, for $1 \leq z \leq k$, co-occurs with $\overrightarrow{b}$ in the same event structures

rated structures, with participants specified for each role.

[6] SMs are akin to Discourse Representation Structures in DRT (Kamp, 2013).

[7] The idea also bears some similarities with the inferential model of communication proposed by Relevance Theory, where the interpretation of a given utterance is the one that maximizes the hearer's expectations of relevance (Sperber and Wilson, 1986).

For instance, the thematic fit of *student* as a subject of *write* is given by the cosine between the vector of *student* and the centroid vector built out of the k most salient subjects of *write*. Similarly, we assess the typicality of *thesis* as an object related to *student* (i.e., as an object of events involving student as subject) by measuring the cosine between the vector of *thesis* and the centoid vector built out of the k most salient objects related to *student*. Finally, we measure in the same way the typicality of *thesis* as an object of *write*.

Formally, the global score θ_{e_k} of an event e_k is defined as:

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

meaning that the degree of semantic coherence of an event is given by the product of the partial thematic fit scores between all its components.[8]

On the other hand, the σ_{e_k} score 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 work, we compute activation of an event e simply by summing the activation scores of the single lexical items cuing it (i.e., the conditional probabilities of the event given each lexical item in the input sentence):

$$\sigma_i = P(e|i) = \frac{P(e,i)}{P(i)} \tag{3}$$

$$F(\sigma_i, \sigma_j) = \sigma_{e_k} = \sigma_i + \sigma_j \tag{4}$$

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

To sum up, we weight 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_L and retrieved during sentence processing. On the the other hand, the θ score allows us to weight events not available in the Memory component. In fact, the Unification component can construct new event never observed before, thereby accounting for the ability to comprehend novel sentences representing atypical and yet possible events. For instance, the event expressed by (1-a) might be expected to be already stored in GEK_L because of its high typicality, thereby having a high σ score. Suppose instead that the sentence (1-b) expresses a brand new event, and that its components never cooccurred together before. In this case, its weight will only depend on the θ score, that is on how similar are its participants to other events stored in the event repository (e.g., how *mechanic* is similar to the prototypical subjects of *write*). Therefore, the joint effect of the σ and θ scores captures the "balance between storage and computation" driving sentence processing (cf. above).

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

$$\text{INT}(s) = \underset{e}{\text{argmax}}(\text{SCW}(e)) \tag{5}$$

$$\text{SCW}(e) = \theta_e + \sigma_e \tag{6}$$

We model the **semantic complexity (SempComp)** of a sentence s as inversely related to the SCW of the event representing its interpretation:

$$\text{SemComp}(s) = \frac{1}{\text{SCW}(\text{INT}(s))} \tag{7}$$

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.

3 Modeling Logical Metonymy

We apply the distributional model of sentence comprehension presented in the previous section to account for psycholinguistic data about metonymic sentences. In particular, we predict that *metonymic sentences will have higher SemComp scores than non-coercion sentences*, because they do not comply with the semantic preferences of the event-selecting verb. According to Zarcone et al. (2013), it is exactly the low thematic fit between verb and object that triggers complement coercion and that, at the same time, causes the extra processing load.

Additionally, we predict that the covert event in metonymic sentence is i.) strongly activated by the lexical items in the context, and is ii.) semantically coherent with respect to the participants that

[8] For the present study, we discarded the modifiers. However, θ scores could also be computed for measuring the coherence of modified arguments (e.g. *the angry child smiled*). We thank one of our reviewers for pointing this out.

are overtly realized. In other words, the inferred covert event is *the event that maximizes the SCW of the global event structure* representing the interpretation of the sentence.

3.1 Datasets

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 compared three different experimental conditions, by contrasting constructions requiring a type-shift with constructions requiring normal composition:

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

Sentence (2-a) corresponds to the metonymic condition (MET), while sentences (2-b) and (2-c) correspond to non-metonymic constructions, with the difference that (2-b) represents a typical event given the subject and the object (HIGH_TYP), whereas (2-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 (33 triplets), while the Traxler dataset was used in the eye-tracking experiment by Traxler et al. (2002) and contains 108 sentences (36 triplets).[9]

3.2 Extracting GEK_L

In order to populate the repository of events in GEK_L, we followed the procedure proposed by 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, 2013). For each sentence, we generated an event (as described in Section 2.1) by extracting the verb and its direct dependencies. In the present case, the dependency relations of interest are subject (SUBJ), 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. We discarded the adjectival/adverbial modifiers and we just keep their heads. For instance, from the joint context *director-n-subj__write-v-head__article-n-dobj* we generated the event $[_{EVENT}$ NSUBJ:$\overrightarrow{student}$ HEAD:$\overrightarrow{read}$ DOBJ:$\overrightarrow{book}]$. For each joint context, we also generated schematic events from its dependency subsets. We totally extracted 1,043,766 events that include 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 the GEK_L. Words with frequency below 100 were excluded. The total number of targets is 20,560 (cf. Table 1 for the dataset coverage). As vector dimensions, we used the same target words, while the dependency relations are the same used to build the joint contexts (*SUBJ:author-n* and *DOBJ:book-n* are examples of dimensions for the target *write-v*). Syntactic co-occurrences were weighted with Local Mutual Information (Evert, 2004):

$$LMI(t,r,f) = log\left(\frac{O_{trf}}{E_{trf}}\right) * O_{trf} \qquad (8)$$

with O_{trf} the co-occurrence frequency of the target t, the syntactic relation r and the filler f, and E_{trf} their expected co-occurrence frequency.

Dataset	Coverage
McElree et al. (2001)	30/33
Traxler et al. (2002)	36/36

Table 1: GEK_L coverage for the evaluation triplets

3.3 Modeling the Processing Cost of Metonymic Sentences

The sentences in the original datasets were represented as S(subject)-V(verb)-O(object) tuples. For each sentence s, SemComp(s) was measured as in equation (7), by computing θ_e and σ_e as follows:

- θ_e is the product of the thematic fit of O given V, $\theta_{O,V}$, the thematic fit of S given V, $\theta_{S,V}$, and the thematic fit of O given S, $\theta_{O,S}$ (see Equation 2). $\theta_{O,V}$ 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 *book* and the centroid vector of the most salient objects of *write*); $\theta_{S,V}$ is the cosine between the vector of S and the centroid vector built out of the

[9] The sentences in the same triple have the same syntactic complexity, as they differ only for the verb.

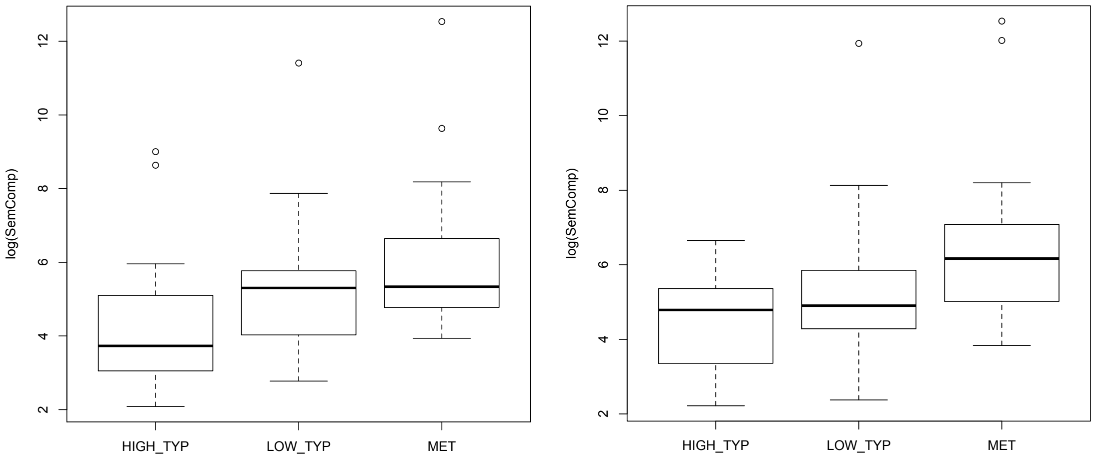

Figure 1: SemComp scores for McElree (left) and Traxler (right)

k most salient subjects of V (e.g., the cosine between the vector of *author* and the centroid vector of the most salient objects of *write*); finally, $\theta_{O,S}$ 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 *book* and the prototype vector of the most salient objects of events whose subject is *author*). Following Baroni and Lenci (2010), we used LMI scores to identify the most salient fillers of each target-specific syntactic slot and we fixed $k = 20$.

- σ_e is the salience score of the triple s, and it corresponds to the sum of the activation scores of i.) the full event represented by the triple and of ii.) the sub-events corresponding to all the partial combinations of the verb and its arguments. Each activation score is the conditional probability of the event given a lexical item in the test tuple.

  Given the verb-argument triple s, the set E is the set of i events containing i.) the entire event e; ii.) all the schematic events $e_1, \ldots, e_i$ generated by abstracting over one of the lexemes of the triples (e.g., for $s = \{author - write - book\}$), $E = \{< author, write, book >, < author, write >, < author, book >, < write, book >\}$. σ_e is computed with the following equation:

$$\sigma_e = \sum_{e_i \in E} \sigma_{e_i} \tag{9}$$

Figure 1 shows the boxplots of the log SemComp scores for three types of sentences (MET, HIGH_TYP, and LOW_TYP) in the datasets. The Kruskal-Wallis rank sum test reveals a main effect of the sentence types on the *SemComp* scores assigned by our GEK_L-based distributional model for the McElree dataset ($\chi^2 = 17.18$, $p < 0.001$). Post-hoc tests (cf. Table 2) show 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).

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

Table 2: Results of the pairwise *post-hoc* comparisons for the three conditions on the McElree dataset (Wilcoxon rank sum test with Bonferroni correction).

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

Table 3: Results of the pairwise *post-hoc* comparisons for the three conditions on the Traxler dataset (Wilcoxon rank sum test with Bonferroni correction).

A main effect of sentence types on the SemComp score also also exists 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 distributional model produces similar results: 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 ($p < 0.05$).

3.4 Identifying the Covert Event

We assume that the interpretation of a metonymic sentence like *The author starts the book* is the following conjunction of events:

(3) $[_{EVENT}\ \text{NSUBJ:}\overrightarrow{author}\ \text{HEAD:}\overrightarrow{start}\ \text{DOBJ:}\overrightarrow{e}]$
$[_{EVENT}\ \text{NSUBJ:}\overrightarrow{author}\ \text{HEAD:}\overrightarrow{e}\ \text{DOBJ:}\overrightarrow{book}]$

where e is the covert event to be recovered (e.g., writing). We modeled covert event retrieval as a binary classification task, as in Zarcone et al. (2012), using the following procedure: i.) 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*); ii.) for each sentence $SV_{met}O$, we computed SCW(e) (cf. equation 6) of the events composing its interpretation, that is $[_{EVENT}$ S V$_{met}$ E$_{cov}]$ and $[_{EVENT}$ S E$_{cov}$ O];[10] iii.) the model **accuracy** was computed as the percentage of test items for which SCW(E$_{cov}$ = HIGH_TYP_EVENT) is higher than SCW(E$_{cov}$ = LOW_TYP_EVENT).

Model	McElree	Traxler
Random	50%	50%
σ	46.66%	30.55%
θ	73.3%	75%
$\sigma + \theta$	80%	77.77%

Table 4: Accuracy of model components and random baseline on the binary classification task for covert event retrieval.

The results for the covert event identification are shown in Table 4. We tested both the full model (SCW = σ + θ) and its σ and θ components separately, to check their contribution to the task. Overall, it can be observed that the full model is the best performing one, classifying correctly just a few items more than the thematic fit-based, θ-only model. Both models are significantly better than the random baseline at $p < 0.05$ on the Traxler dataset, whereas only the full model achieves a significant advantage over the baseline on McElree.[11]

The performance of the σ component, which makes use only of the information stored in GEK_L, is pretty weak, especially on the Traxler dataset. This is the same problem affecting purely probabilistic approaches, given also the fact that many of the words of the evaluation datasets have low frequencies in corpora. The θ component therefore plays a crucial role in the covert event prediction. In fact, θ works like a generalization component, and it serves to compute and weight new event representations when the information stored in memory is not sufficient. The strong performance of a thematic fit-based method is also consistent with the results obtained by Zarcone et al. (2012) on German data.

Interestingly, a further study by Zarcone et al. (2013) has proposed thematic fit estimation as the mechanism which is responsible also for the triggering of logical metonymy, hypothesizing that the recovery of the implicit event could be a consequence of the dispreference of the verb for the entity-denoting argument. This means, in our perspective, 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 situation model. To test this claim, we compared the θ scores of the events containing the HIGH_TYP covert event (i.e., $[_{EVENT}$ S V$_{met}$ E$_{cov}]$ + $[_{EVENT}$ S E$_{cov}$ O]) and the corresponding MET event (i.e., $[_{EVENT}$ S V$_{met}$ O]), predicting that the former events are more semantically coherent than the latter.[12] 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 events containing the covert events are significantly higher.

4 Conclusions

In this paper, we have presented a distributional model of sentence comprehension as an incremental process to build the semantic representation of the event expressed by the sentence. Events are represented with complex formal structures that contain the distributional vectors of its component. Sentence interpretation is carried out by unifying stored distributional information about

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

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

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

events, GEK_L. The event representing a sentence is the event with the highest semantic composition weight, SCW, which is in turn a function of its internal semantic coherence and the activation strength by the linguistic input. The semantic coherence of an event, measured by the θ score, depends on its similarity to stored events. Therefore, the unlimited ability of understanding new sentences can be conceived as the ability to adapt our general knowledge about events to novel situations: in brief, **productivity is adaptation**, and **adaptation is by similarity**.

The model has been successfully applied to the case of logical metonymy, accounting for two aspects of this phenomenon that have always been treated separately in the literature, namely processing costs and covert event retrieval. Given these encouraging results, we are planning to apply the model also to other semantic tasks involving event knowledge, such as the detection of anomalies (e.g. violations of selectional restrictions), the recovery of implicit arguments and of bridging inferences.

Acknowledgments

This work has been carried out thanks to the support of the A*MIDEX grant (nANR-11-IDEX-0001-02) funded by the French Government 'Investissements d'Avenir' program. We would like to thank the three anonymous reviewers for their many insightful comments and suggestions.

References

Gerry T. M. Altmann. 1999. Thematic Role Assignment in Context. *Journal of Memory and Language* 41(1):124–145.

Nicholas Asher. 2015. Types, Meanings and Coercions in Lexical Semantics. *Lingua* 157:66–82.

Giosuè Baggio and Peter Hagoort. 2011. The Balance between Memory and Unification in Semantics: A Dynamic Account of the N400. *Language and Cognitive Processes* 26(9):1338–1367.

Giosuè Baggio, Michiel van Lambalgen, and Peter Hagoort. 2012. The Processing Consequences of Compositionality. In Markus Werning, Wolfram Hinzen, and Edouard Machery, editors, *The Oxford Handbook of Compositionality*, Oxford University Press, Oxford, pages 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* 43(3):209–226.

Marco Baroni and Alessandro Lenci. 2010. Distributional Memory: A General Framework for Corpus-based Semantics. *Comput. Linguist.* 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* 63:489–505.

Robyn Carston. 2002. *Thoughts and Utterances*. Blackwell, Oxford.

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

Emmanuele Chersoni, Alessandro Lenci, Enrico Santus, Philippe Blache, and Chu-Ren Huang. 2016b. Representing Verbs with Rich Contexts: An Evaluation on Verb Similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 1967–1972.

Roberto G. de Almeida. 2004. The Effect of Context on the Processing of Type-shifting Verbs. *Brain and Language* 90:249–261.

Roberto G. de Almeida and Veena D. Dwivedi. 2008. Coercion without Lexical Decomposition: Type-shifting Effects Revisited. *Canadian Journal of Linguistics* 53(2/3):301–326.

Jeffrey L. Elman. 2009. On the Meaning of Words and Dinosaur Bones: Lexical Knowledge without a Lexicon. *Cognitive Science* 33(4):547–582.

Jeffrey L. Elman. 2014. Systematicity in the Lexicon: On Having your Cake and Eating it too. In Paco Calvo and John Symons, editors, *The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*, The MIT Press, Cambridge, MA.

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

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis.

Ingrid L. Falkum. 2011. A Pragmatic Account of Logical Metonymy'. In *Proceedings of Metonymy 2011*. pages 11–17.

Steven Frisson and Brian McElree. 2008. Complement Coercion is not Modulated by Competition: Evidence from Eye Movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34(1):1–11.

Peter Hagoort. 2013. MUC (Memory, Unification, Control) and Beyond. *Frontiers in Psychology* 4(JUL):1–13.

Peter Hagoort. 2016. MUC (Memory, Unification, Control): A Model on the Neurobiology of Language beyond Single Word Processing. In Gregory Hickok and Steve Small, editors, *Neurobiology of Language*, Elsevier, Amsterdam, volume 28, pages 339–347.

Mary Hare, Michael N. Jones, Caroline Thomson, Sarah Kelly, and Ken McRae. 2009. Activating Event Knowledge. *Cognition* 111:151–167.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. The MIT Press, Cambridge, MA.

Yuki Kamide, Gerry T. M. Altmann, and Sarah L. Haywood. 2003. The Time-course of Prediction in Incremental Sentence Processing: Evidence from Anticipatory Eye Movements. *Journal of Memory and Language* 49:133–156.

Hans Kamp. 2013. *Meaning and the Dynamics of Interpretation: Selected Papers by Hans Kamp*. Brill, Leiden-Boston.

Gina R. Kuperberg. 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.

Mirella Lapata, Frank Keller, and Christoph Scheepers. 2003. Intra-sentential Context Effects on the Interpretation of Logical Metonymy. *Cognitive Science* 27(4):649–668.

Mirella Lapata and Alex Lascarides. 2003. A Probabilistic Account of Logical Metonymy. *Computational Linguistics* 29(2):261–315.

Alex Lascarides and Ann Copestake. 1998. Pragmatics and Word Meaning. *Journal of Linguistics* 34:378–414.

Geoffrey Neil Leech. 2013. 100 Million Words of English: the British National Corpus (bnc)*.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research* 5(Apr):361–397.

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

Ken McRae, Mary Hare, Jeffrey L. Elman, and Todd R. Ferretti. 2005. A Basis for Generating Expectancies for Verbs from Nouns. *Memory & cognition* 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* 3(6):1417–1429.

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

James Pustejovsky. 1995. *The Generative Lexicon*. The MIT Press, Cambridge, MA.

Ekaterina Shutova. 2009. Sense-based Interpretation of Logical Metonymy Using a Statistical Method. In *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*. pages 1–9.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. Blackwell, Oxford.

Matthew J. Traxler, Martin J. Pickering, and Brian McElree. 2002. Coercion in Sentence Processing: Evidence from Eye-movements and Self-paced Reading. *Journal of Memory and Language* 47:530–547.

Alessandra Zarcone, Alessandro Lenci, Sebastian Padó, and Jason Utt. 2013. Fitting, not Clashing! A Distributional Semantic Model of Logical Metonymy. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*. pages 404–410.

Alessandra Zarcone and Sebastian Padó. 2011. Generalized Event Knowledge in Logical Metonymy Resolution. In *Proceedings of the 33rd annual meeting of the Cognitive Science Society (CogSci 2011)*.

Alessandra Zarcone, Sebastian Padó, and Alessandro Lenci. 2014. Logical Metonymy Resolution in a Words-as-Cues Framework: Evidence from Self-paced Reading and Probe Recognition. *Cognitive Science* 38(5):973–996.

Alessandra Zarcone, Jason Utt, and Sebastian Padó. 2012. Modeling Covert Event Retrieval in Logical Metonymy: Probabilistic and Distributional Accounts. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*. pages 70–79.

Rolf A. Zwaan and Gabriel A. Radvansky. 1998. Situation Models in Language Comprehension and Memory. *Psychological bulletin* 123(2):162–185.