AIX-MARSEILLE UNIVERSITÉ

UNIVERSITÉ DE PISE

ED 356 Cognition, Langage, Éducation

LPL - Laboratoire Parole et Langage

Thèse présentée pour obtenir le grade
universitaire de Docteur en Sciences du Langage

Emmanuele Chersoni

Explaining Complexity in Human Language Processing : A Distributional Semantic Model

Directeur de thèse : Prof. Philippe BLACHE

Co-directeur de thèse : Prof. Alessandro LENCI

Soutenue le 04/07/2018 devant le jury composé de :

Vera DEMBERG	Université de la Sarre	Rapporteur
Vito PIRRELLI	ILC-CNR Pise	Examinateur
Laurent PRÉVOT	Aix-Marseille Université & LPL-CNRS	Examinateur
Tim VAN DE CRUYS	IRIT-CNRS	Examinateur
Roberto ZAMPARELLI	Université de Trente	Rapporteur
Philippe BLACHE	Aix-Marseille Université & LPL-CNRS	Directeur de thèse
Alessandro LENCI	Université de Pise	Co-directeur de thèse

Numéro national de thèse/suffixe local : 2017AIXM0001/001ED62

Table des matières

Résumé

Le présent travail aborde le thème de la complexité sémantique dans le langage naturel, et il propose une hypothèse basée sur certaines caractéristiques des phrases du langage naturel qui déterminent la difficulté pour l'interpretation humaine.
Nous visons à introduire un cadre théorique général de la complexité sémantique de la phrase, dans lequel la difficulté d'élaboration est liée à l'interaction entre deux composants : la Mémoire, qui est responsable du rangement des représentations d'événements extraites par des corpus, et l'Unification, qui est responsable de la combinaison de ces unités dans des structures plus complexes. Nous proposons que la complexité sémantique depend de la difficulté de construire une représentation sémantique de l'événement ou de la situation exprimée par une phrase, qui peut être récupérée directement de la mémoire sémantique ou construit dynamiquement en satisfaisant les contraintes contenus dans les constructions.
Pour tester nos intuitions, nous avons construit un *Distributional Semantic Model* pour calculer le coût de composition de l'unification des phrases. Les tests sur des bases de données psycholinguistiques ont révélé que le modèle est capable d'expliquer des phénomène sémantiques comme la mise à jour *context-sensitive* des attentes sur les arguments et les métonymies logiques.

Mots clés : sémantique distributionelle, complexité sémantique, compréhension de la phrase, N400, memoire sémantique, unification, psycholinguistique computationelle, *thematic fit*

Abstract

The present work deals with the problem of the semantic complexity in natural language, proposing an hypothesis based on some features of natural language sentences that determine their difficulty for human understanding.

We aim at introducing a general framework for semantic complexity, in which the processing difficulty depends on the interaction between two components: a Memory component, which is responsible for the storage of corpus-extracted *event representations*, and a Unification component, which is responsible for combining the units stored in Memory into more complex structures. We propose that semantic complexity depends on the difficulty of building a semantic representation of the event or the situation conveyed by a sentence, that can be either retrieved directly from the semantic memory or built dynamically by solving the constraints included in the stored representations.

In order to test our intuitions, we built a *Distributional Semantic Model* to compute a *compositional cost* for the sentence unification process. Our tests on several psycholinguistic datasets showed that our model is able to account for semantic phenomena such as the context-sensitive update of argument expectations and of logical metonymies.

Keywords: distributional semantics, semantic complexity, sentence comprehension, N400, semantic memory, unification, thematic fit, computational psycholinguistics

Acknowledgements

I would like to thank sincerely my supervisors, Prof. Philippe Blache and Prof. Alessandro Lenci, for their continuous support of my research, for the patience and motivation that they have shown in guiding me in the last four years, and for having been a constant source of feedback and ideas during my studies.

Besides my advisors, my sincere thanks also go to my friend and colleague Enrico Santus, who has been my collaborator for many of the scientific investigations presented here: his knowledge and his experience have been an essential reference point for me to improve as a researcher, and without his help many of the goals of the present work would not have been achieved.

Last but not least, I would also like to thank my family, for supporting me spiritually during these years and for making the values of work and commitment a part of my education.

Introduction: What Kind of Linguistic Complexity?

The study of *linguistic complexity* has a longstanding tradition in theoretical and computational linguistics, and over the years it has attracted several contributions from different research fields. On the one hand, understanding complexity is a fundamental issue for studying the mechanisms behind human language comprehension, given its close connections with the nature of the operations involved in building representations for the sentences that we hear or read in our everyday life. On the other hand, the topic is particularly appealing also for its potential applications, such as readability assessment and automatic text simplification.

The primary question related to complexity concerns the identification of *complexity factors*, namely what makes natural language sentences more or less difficult to process. Linguistic theory has essentially provided two types of answers (Blache 2011; Brunato 2015):

— in a typological perspective, complexity depends on the formal properties of the language system, and it is measured by means of comparisons between different languages. Complexity factors are conceived as residing at the competence level, and different degrees of complexity between languages are generally explained in terms of the number of (morphological, syntactic, semantic etc.) distinctions that a speaker has to master (McWhorter 2001; Parkvall 2008);

— alternatively, complexity can be a property of concrete linguistic realizations (sentences, utterances etc.), and it can be described as the

processing difficulty experienced by language users (Blache 2011; Rauzy and Blache 2012). Such a perspective, which looks for the complexity factors at the performance level, has been mainly adopted by psycholinguistic and neurolinguistic studies.

The present study adopts the second perspective, and deals with the issue of the sources of complexity in sentence processing, sticking with the evidence that has been brought by experimental studies of the last two decades and trying to bring it together into a unitary theoretical framework. More specifically, the research results we have been focusing on mainly concern the differences in processing between sentences that differ for some semantic factors. By *differences in processing*, we simply mean that some sentences are easier for human understanding (i.e. they require less cognitive effort) than others. From this point of view, the notions of linguistic *complexity* and linguistic *difficulty* in our framework are overlapping. We believe that accounting for such differences is necessary to shed new light on the organization of the semantic memory and on the process of construction of meaning during sentence comprehension.

The second main goal of the present research is to propose an implementation of our theoretical proposal by means of a Distributional Semantic Model (Lenci 2008; Turney and Pantel 2010), whose predictions will be compared with the outcomes of some of the above-mentioned studies on processing complexity.

0.1. Psycholinguistic Perspectives on Complexity

Our study is certainly not the first to try to identify parameters of linguistic complexity and to evaluate a computational model with respect to psycholinguistic evidence. An example of a well-known account of linguistic complexity that was designed with this specific purpose is Ted Gibson's Dependency Locality Theory (DLT), originally described in Gibson 1998 and Gibson 2000. In DLT, complexity mainly depends on the number of referential objects occurring between two syntactic structures. A second factor is the novelty of the discourse referents (typically introduced by a pronoun or by a proper noun), which is assumed to determine extra processing difficulties.

More concretely, Gibson's theory distinguishes between two types of processing costs:

— *integration costs*, defined as the distance between a head and its governor and computed as the number of new discourse referents between the two;

— *storage costs*, the minimal number of syntactic heads that are necessary in order to form a grammatical sentence (e.g. after a subject, a verb is required to complete the sentence).

Consider the following example:

(1) a. The reporter disliked the editor.

b. The reporter [who the senator attacked] disliked the editor.

According to DLT, sentences like (1b) are more difficult to process as there

is an embedded relative clause between the subject and the predicate, and a new discourse referent is introduced in it. Beyond its influence on psycholinguistics-oriented studies on complexity, DLT can also be seen as bridge between the two perspectives on the problem, given the recent evidence for the hypothesis that the grammars of all languages tend to minimize dependency length (Gildea and Temperley 2010; Futrell, Mahowald, and Gibson 2015).

Following Gibson's work, other approaches have proposed to take into account also the *activation* of linguistic structures as a factor in the quantification of complexity (Vasishth 2003a; Vasishth 2003b; R. L. Lewis and Vasishth 2005). When words or grammatical categories are particularly predictable given a context, the integration in a structure should be easier. For example, Vasishth 2003a showed that sentences like (2b) are processed more quickly than those in (2a), contrary to the DLT prediction.

(2) a. The rat the cat saw died.
b. The rat the cat briefly saw died.

In other words, the presence in (2b) of the adverbial pre-modifier works as a facilitating factor, as it increases the activation of the verbal head.

An alternative approach adopts **surprisal** as a metric for sentence complexity (Hale 2001; Hale 2016). Such a notion is grounded on a probabilistic view on language and corresponds, roughly speaking, to the probability mass of the analyses that are not consistent with a new incoming word (R. Levy 2008; Demberg and Keller 2008). Therefore, it is possible to estimate an index of sentential complexity simply by summing the surprisal values associated to the single words.

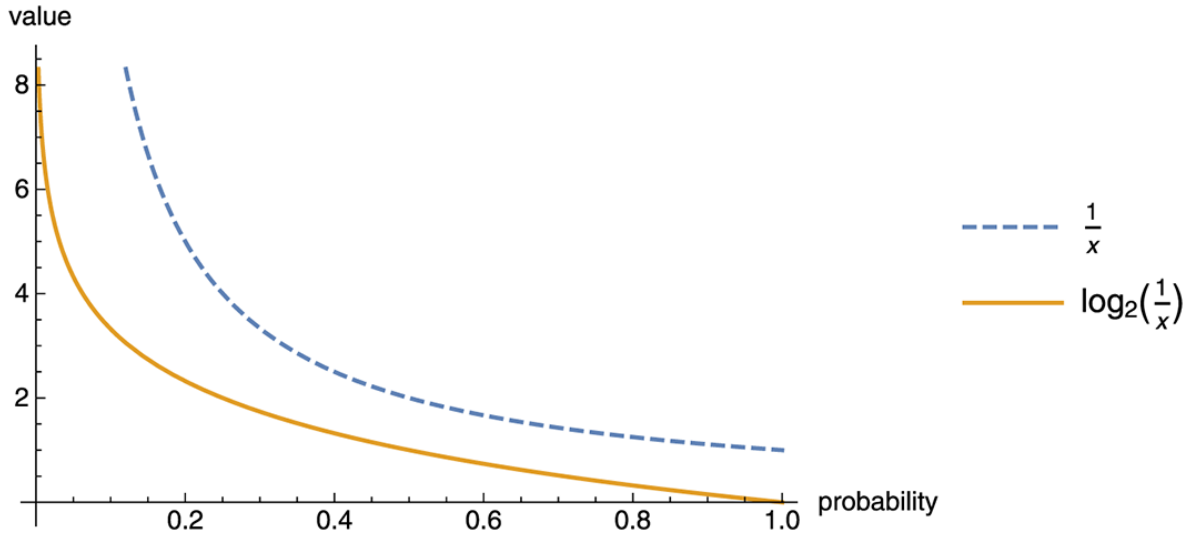


Figure 0.1. – Given a random variable Y, the surprisal of the outcome Y = y is equal to $\log \frac{1}{P(y)}$. As probabilities get closer to zero, surprisal values get higher (image taken from Hale 2016).

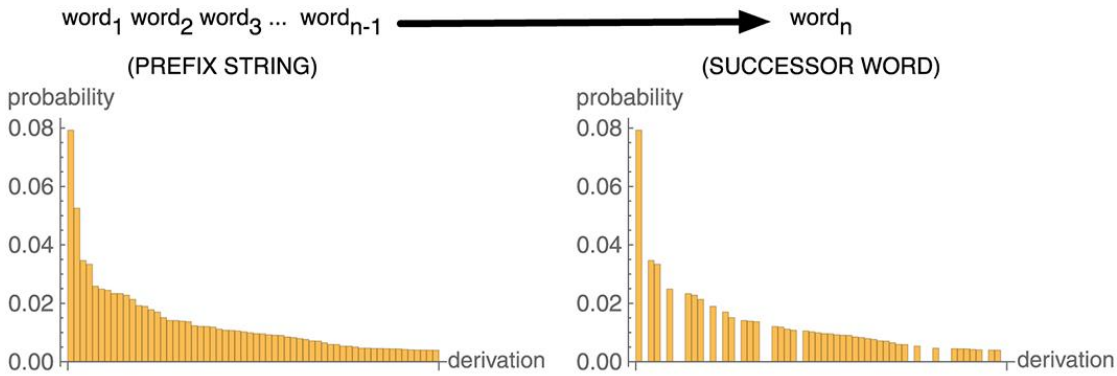


Figure 0.2. – The image from Hale 2016 shows a concrete example of the idea of surprisal to the grammatical derivations in a probabilistic grammar. On the left, the histogram represents the probability distribution over the grammatical derivations after a given prefix string. When the next word ($word_n$) appears, some of the derivations are ruled out, and the same goes for the next transitions until one derivation is left. The probability of a transition is defined as a ratio of probability sums, $\frac{afterTransition}{beforeTransition}$, and the surprisal of the new word is equivalent to the reciprocal of this ratio. In other words, the higher the probability mass that is ruled out as the new word appears, the higher its surprisal value.

When tested on eye-tracking data, complexity models based on surprisal

have been proven to be able to predict sentence reading times (Demberg and Keller 2008; Rauzy and Blache 2012) and word pronunciation duration (A. Sayeed, Fischer, and Demberg 2015), which is also considered as an indicator of complexity. Moreover, the approach is relatively easy to integrate with non-syntactic factors, e.g. the semantic relatedness of the target word with other words in the previous linguistic context (Pynte, New, and Kennedy 2008).

0.2. Thesis Overview

The short resume that we just presented should be sufficient to understand that research on linguistic complexity has especially focused on syntactic factors, eventually enriching the syntactic models with some extra features related to other linguistic domains (e.g. lexical semantics).

In this study, we adopt a different perspective, as we are specifically focusing on semantic factors of difficulty. Moreover, we also present a hypothesis on the type of representations that are stored in the semantic memory and how they can determine differences in processing complexity, with reference to the results in the experimental literature.

The thesis is organized as follows:

— in the first chapter, we present an overview of the recent advances in the experimental linguistics literature on sentence processing that motivate our proposal. A specific focus has been given to i) experimental evidence for the early access to event knowledge during online sentence processing; ii) psycholinguistic models that re-frame the sentence comprehension problem as a *cue-to-category mapping*;

— in the second chapter, we introduce the computational and method-

ological framework that we used to set up our model, i.e. Distributional Semantic Models (DSMs). A particular attention will be given to distributional models aiming at modeling event knowledge;

— in the third chapter, we provide an overview of our model of processing complexity, as well as a proposal for a computational implementation. Some case studies from the psycholinguistic literature will be also introduced, in order to show the ability of the model to account for different semantic phenomena;

— in the last and conclusive chapter, we discuss the future perspectives of our research and the potential improvements to our complexity model.

1. Background

Main contributions of the chapter:

— the problem of compositionality in language processing is introduced in the perspective of the Memory, Unification and Control framework proposed by Peter Hagoort;

— we present the experimental evidence for a semantic memory containing knowledge about events and their typical participants;

— we present an overview of models that consider the comprehension process as a cue-to-category mapping problem.

1.1. The Compositionality Problem and the Content of Semantic Memory

Generally cited as a strong argument in favor of the principle of compositionality (Pelletier 1994), the human capacity of producing and understanding novel and meaningful sentences has recently been at the core of an intense debate in experimental linguistics.

Let us consider the following examples:

(1) a. *The musician plays the flute in the theater.*

b. *The gardener plays the castanets in the cave.*

c. ** The nominative plays the global map in the pot.*

Since the early work of Chomsky (Chomsky 1957) and the introduction of the notion of the acceptability of a sentence, the traditional focus of linguistic theory was the contrast between sentences like (1a-1b) and sentences

like (1c).

The last sentence clearly differs from the first two as it is semantically impossible, albeit syntactically well-formed. In other words, (1c) violates the combinatorial constraints of the lexical items at such a degree that we are not able to build any coherent representation for the described situation. There is also a striking difference between (1a) and (1b): although both the sentences are interpretable (i.e. we can build a semantic representation of what is going on), the event expressed by (1b) sounds weird and unusual.

The human ability of distinguishing between the second and the third example can be described as the capacity to distinguish between novel and meaningful sentences (1b) and semantically impossible ones (1c). Moreover, the set of novel sentences that we can produce and understand is apparently open-ended. For traditional linguistic theories, the goal of characterizing semantic compositionality entails, by consequence, identifying the rules and the principles according to which lexical items can combine to produce meaningful combinations, as well as the way in which their semantic content can be represented.

However, recent experimental researches on Event-Related Potentials (henceforth ERP) and language processing [1] have shifted the focus on a different aspect of compositionality, namely the fact that typical and familiar sentences like (1a) might have a different cognitive status (i.e. they are processed in a different way) from sentences like (1b), which are made up

1. The Event Related Potentials technique, in cognitive neuroscience, is used to record large-scale electrical activity in the brain, by means of electrodes that are placed on the subject's scalp. Such a technique has become increasingly popular in the studies on online sentence processing, especially because of its high temporal resolution that allows to match the fast rate of language comprehension and to detect the brain response to some linguistic stimulus (see Kaan 2007 for a more detailed overview).

of *possible, but unexpected lexical combinations*. In particular, the latter type of sentences evoke *stronger N400 components* in the ERP waveform than sentences including just *predictable and familiar combinations* (Baggio and Hagoort 2011). [2]

Two major interpretations of the N400 effect have been proposed by researchers:

— according to the 'feature pre-activation hypothesis' [3], the sentence and the discourse context pre-activate the semantic features of the upcoming lexical items (Federmeier and Kutas 1999; Kutas and Federmeier 2000; Van Berkum, C. M. Brown, Zwitserlood, Kooijman, and Hagoort 2005). The more an upcoming word is coherent with the expectations generated from the context (i.e. it matches the activated lexical features), the smaller the elicited N400 will be; [4]

— according to the 'semantic unification hypothesis', the N400 amplitude is proportional to the difficulty in integrating the meaning of a current word with the current representation of the contextual meaning (Baggio and Hagoort 2011; Baggio, Van Lambalgen, and Hagoort 2012). Words that are difficult to integrate with the context will generate, as a processing consequence, larger N400 amplitudes (see Figure 1.1).

2. As the ERP is a sequence of positive- and negative-going deflection, several waveforms (also called *components*) have been identified (primarily) on the basis of their i) polarity and ii) timing of the onset or of the peak. An important matter of speculation in the field concerns the underlying functional processes (for example, syntactic/semantic unification) that might be reflected by a specific component (Kaan 2007).

3. The names of both the hypotheses were not proposed by the cited authors: they are introduced here for indicative purposes only.

4. In Rabovsky and McRae 2014, a similar idea is conceptualized in information-theoretic terms, as the N400 reflects the *surprisal* at the level of semantic features, i.e the amount of semantic information of the incoming stimulus that has not been prediced by the context.

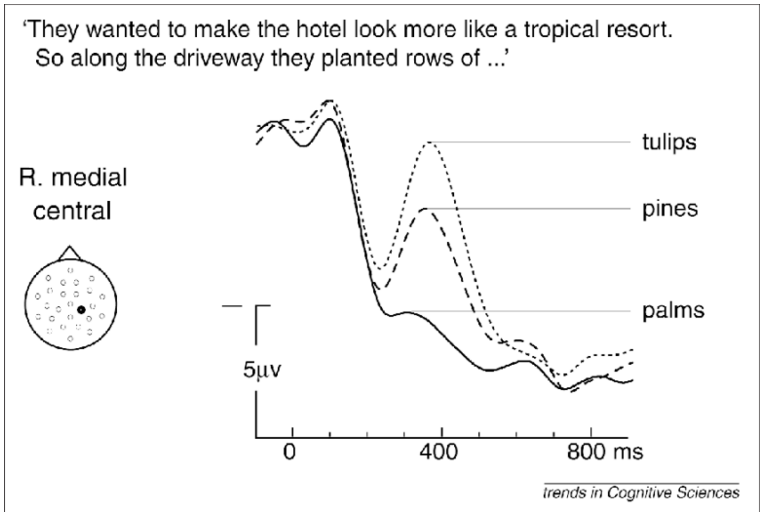


Figure 1.1. – An image taken from Kutas and Federmeier 2000, showing the relationship between the contextual predictability of a word and the amplitude of the N400 component. As *palms* is the most predictable word in the given context, it elicits a small N400, whereas the component amplitude is increasingly larger for unexpected words (larger for *tulips* because, unlike *pines*, it does not belong to the same basic-level semantic category of the expected word, i.e. *plants*).

It is evident that both interpretations agree on the fact that novel combinations of lexemes require a bigger cognitive effort to be understood: in the first case because they are unexpected, in the second case because they are not coherent with the unfolding semantic representation of a wider context.

A fundamental claim of Baggio, Van Lambalgen, and Hagoort 2012 is that the real issue about compositionality and open-ended productivity in natural language is 'the balance between storage and computation'. On the one hand, productivity entails that we cannot store every possible utterance: there must be a computational mechanism that allows us to combine stored units in order to generate representations for new and unseen combinations.

On the other hand, the processing differences showed by the N400 ef-

fects, together with the evidence brought by a long trend of psycholinguistic studies (see McRae and Matsuki 2009 for an overview), suggest that our semantic memory stores a large amount of knowledge about event contingencies and specific combinations of concepts. This 'event knowledge' is activated during sentence processing and potentially acts as a facilitating factor on unification: the more the situation described by the sentence is coherent with the content of the semantic memory, the easier it will be to create a unified representation for it, and the more the processing effort will be reduced.

In the field of neurosciences, such a division of labor between storage and computation has been accounted for by a general architecture of the neurobiology of language, i.e. the Memory, Unification and Control model (MUC) proposed by Peter Hagoort (Hagoort 2013; Hagoort 2016). Hagoort's model includes the following main components:

— **Memory** refers to the linguistic knowledge stored in the long-term memory and is defined as the only language-specific component. The units stored in Memory are unification-ready structures with different degrees of internal complexity (morphemes, words, phrasal idioms, syntactic frames are all examples of structures that can be stored) and bear a resemblance with the **constructions** described by constructionist theories of grammar (A. E. Goldberg 1995; A. E. Goldberg 2006; Sag 2012; Bybee 2013). Constructions are represented by sets of **constraints** pertaining to the various levels of linguistic representation (phonology, syntax, semantics etc.). The constraints specify how each constructions can combine with the others, as well as the output of such a unification process;

— **Unification** refers to the mechanism assembling the units stored in memory into larger and more complex structures, with contributions from both the linguistic and the extralinguistic context. Unification is a constraint-based process, aiming at the satisfaction of the constraints of the constructions to be unified, and unification operations take place in parallel at all the representation levels;

— **Control** is the component responsible for relating language to joint action and social interaction and it is involved, for example, for all the aspects of the usage of language in conversational settings.

MUC's 'closest relative' in the linguistic theory is probably the Parallel Architecture framework proposed by Ray Jackendoff (Jackendoff 1997; Jackendoff 2003; Culicover and Jackendoff 2005): the theories have in common a) the idea of a lexicon conceived as a repository of 'constructions' stored in the long-term memory; b) their different degrees of internal complexity and schematicity; c) their definition in terms of multi-level constraints; d) unification as a constraint-driven process carried out in parallel at different representation levels.

A important issue concerning both models is: what type of information, precisely, is stored in the long-term memory? In particular, how is it acquired and from which sources? A common *tenet* of modular theories of sentence processing is that the comprehension system initially uses just the knowledge available within specialized modules, and only later uses world knowledge (Frazier 1987; Warren, Milburn, Patson, and Dickey 2015). In the MUC model, for example, the two types of information interact only at the Unification level, as the Unification operations integrate both sorts of knowledge simultaneously (Hagoort 2016).

However, the present work adopts a different point of view on the content of the Memory component. As anticipated before, we believe that there is strong evidence in favor of the idea of a semantic memory containing a **Generalized Event Knowledge**, and against the hypothesis of a modular lexicon.

Therefore, the following section will be dedicated to the presentation of the experimental studies supporting this claim.

1.2. Generalized Event Knowledge in the Experimental Research

At the end of the 90s, one of the most popular framework for sentence processing was the *garden-path model* (also known as the *two-stage model*) developed by Frazier and colleagues (Frazier 1987), which belongs to the family of the modular theories. Indeed, the primacy of syntax is a main assumption of such a framework: the syntactic information is used to assign an initial structure to the linguistic input (i.e. the syntactic representation is constructed *without consulting non-syntactic information sources*; see Ferreira and Clifton 1986; Van Gompel, Pickering, Pearson, and Liversedge 2005; Clifton and Staub 2008), while semantic information becomes available only at a later stage and can be used for eventual revisions of the initial structure.

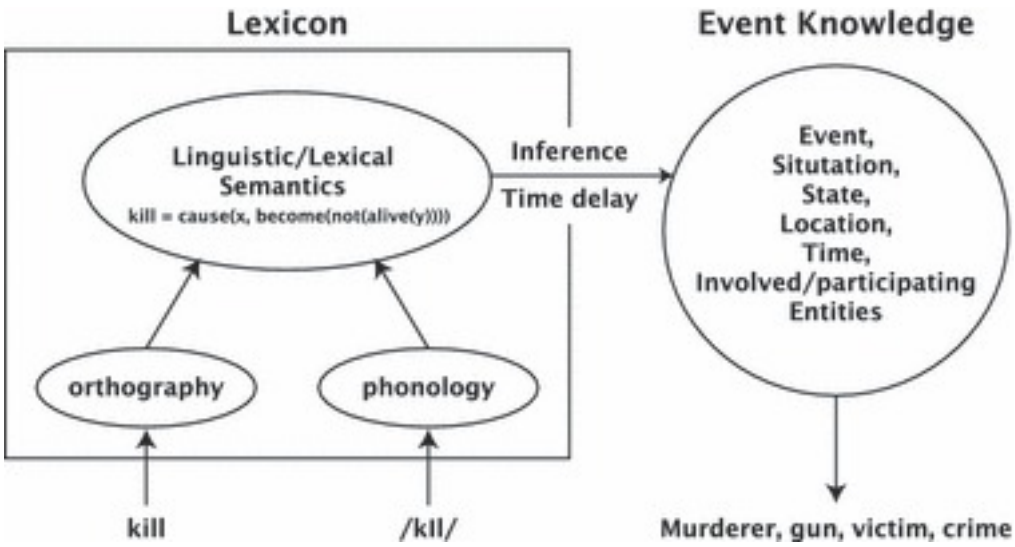


Figure 1.2. – An image taken from McRae and Matsuki 2009, illustrating the separation between linguistic and world knowledge in modular theories.

A modular view of the lexicon was the most natural complement for such theories: linguistic information residing in the lexicon was conceived as clearly distinct from pragmatic and world knowledge. When people encounter a word, consequently, only its 'core', lexical meaning gets activated, while the use of general world knowledge is delayed.

In both cases, it is assumed that some 'linguistically-relevant' type of knowledge that is supposed to be used initially in the comprehension process, with the impact of non-linguistic sources of knowledge coming only after a delay *per se* (Jackendoff 2003; Nieuwland and Van Berkum 2006; see Figure 1.2).

In contrast with the modular view, constraint-based models of language processing assume that we build simultaneously more alternative analysis of the syntactic structure of a sentence, and that the comprehension system quickly exploits all the available sources of information (even the

non-linguistic ones) to identify the analysis that best satisfies the set of the contextual constraints. [5]

Within the research area of constraint-based models, McRae, M. Spivey-Knowlton, and Tanenhaus 1998 dedicated a study to the influence of the so-called *thematic fit* constraint on the selection between two possible syntactic interpretations of a sentence, i.e. the Main Clause vs. the Reduced Relative interpretation. This work inherited from the Competition and Integration model by M. J. Spivey-Knowlton 1996 an essential feature of constraint-based frameworks, namely the idea of having multiple sources of information represented as independent constraints that provide probabilistic support for competing syntactic analyses (see Figure 1.3).

In the experimental data of McRae, M. Spivey-Knowlton, and Tanenhaus 1998, the syntactic ambiguity was encountered immediately after the verb in sentences such as *The man arrested...*, which are open to two possible continuations: one in which *arrested* is the main verb in an active sentence, and the other one in which it is a past participle, followed by a *by*-introduced prepositional phrase and within a reduced relative clause.

5. See McRae and Matsuki 2013 for a complete overview on constraint-based models.

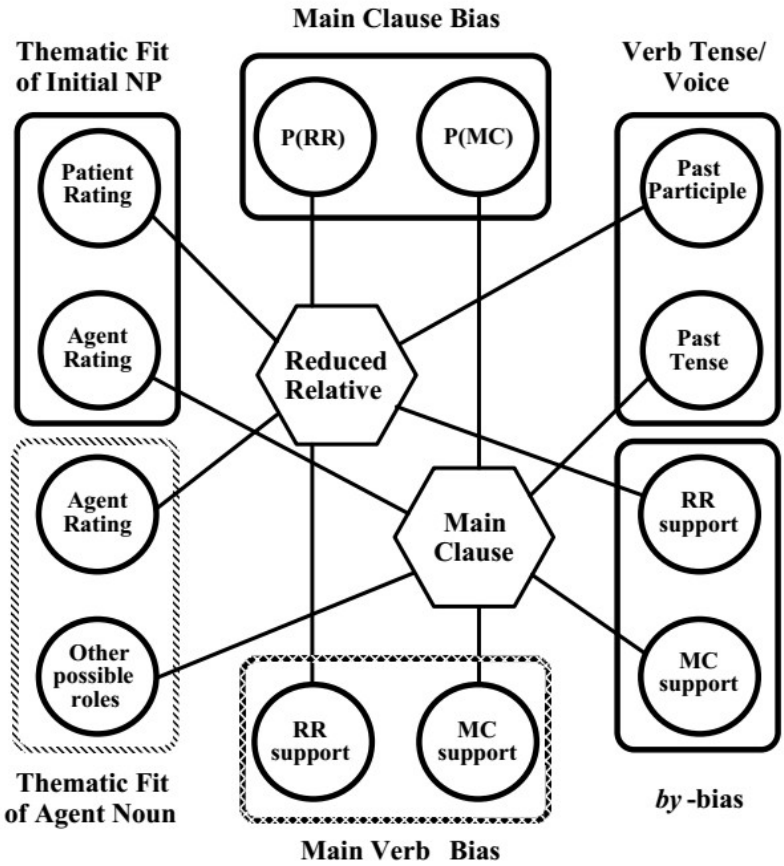


Figure 1.3. – An image taken from McRae, M. Spivey-Knowlton, and Tanenhaus 1998. The *Main Clause Bias*, the *Verb Tense/Voice*, the *by-bias*, the *Main Verb Bias*, the *Thematic Fit of the Agent Noun* and the *Thematic Fit of Initial NP* are the constraints providing support for either the Main Clause or to the Reduced Relative interpretation.

One of the main findings of McRae and colleagues was that an important role was played by the goodness of fit of the initial noun either as an agent or as a patient of the following verb: sentences whose initial noun phrase was a typical *agent* of the verb (e.g. *The cop arrested...*) were easier to understand when they continued as active clauses (*The cop arrested the thief, who had tried to escape*), whereas sentences beginning with typical *patients* (e.g. *The thief arrested...*) were more easily understood when they continued as reduced relative clauses (*The thief arrested by the cop had*

tried to escape). Moreover, the authors contrasted the predictions of two versions of the model: one in which the model used immediately all the constraints, as soon as the words relevant for each constraints had been processed, and one in which thematic fit information was delayed, in order to simulate the account of a two-stage processing model. They found that only the former version of the model was able to provide a good fit to the reading time data of the experimental subjects.

Such results showed that, during sentence comprehension, lexical items activate a general knowledge about events and their typical participants, and that such knowledge was quickly used by the experimental subjects to update the unfolding representation of the sentence being processed.

It was subsequently shown that the activation of event knowledge through the lexical items is not limited to sentential or discourse contexts: T. R. Ferretti, McRae, and Hatherell 2001 used the short Stimulus-Onset Asynchrony priming paradigm (SOA) [6] to test if typical verb role fillers were activated also by the verb in isolation.
The answer was affirmative, as the verbs primed the typical fillers of their thematic roles, including:

— typical agents (*to arrest* primes *cop*);

— typical patients (*to arrest-crook*);

— typical instruments (*to stir-spoon*).

6. By *prime*, we mean a word that can be either semantically related or not to another *target word*, which is the stimulus of our experiment. *Semantic priming paradigms* analyze the increases in speed/accuracy between the responses to target words preceded by related primes and those preceded by neutral primes. The processing advantage caused by a related prime in the former condition is called *priming effect*, and studies on lexical semantic relatedness showed that it can be found even with a very short time interval between the presentations of primes and targets (short stimulus-onset asynchrony; see for example Küper and Heil 2009; Hare, M. Jones, Thomson, Kelly, and McRae 2009).
See also Neely 1991 for an overview on priming paradigms.

Moreover, McRae, Hare, Elman, and T. Ferretti 2005 showed that event-based priming works also in the opposite direction: entity-denoting nouns were found to prime verbs referring to the events they typically participate in, and together with the above-mentioned thematic roles, robust effects were obtained also for *locations*.

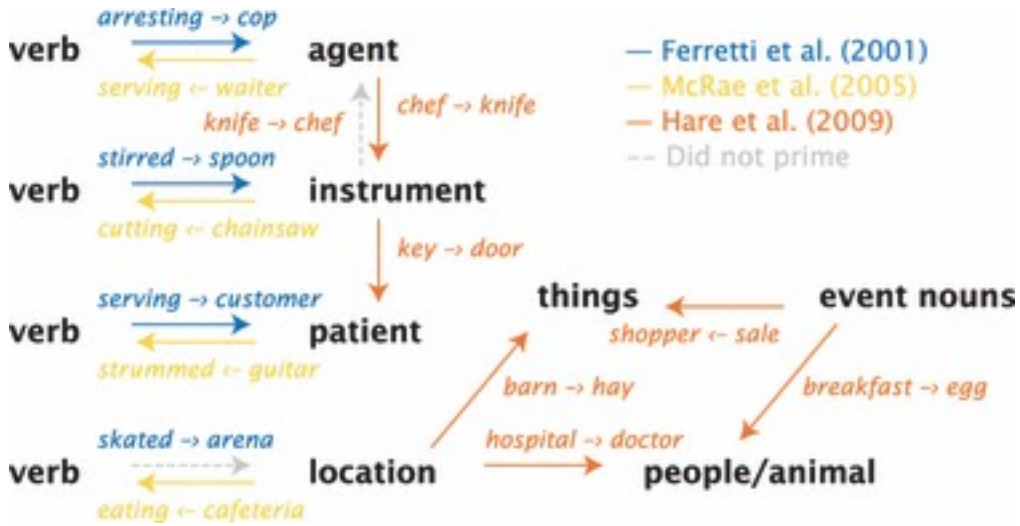


Figure 1.4. – An image taken from McRae and Matsuki 2009, summarizing the experimental studies on event-based verb-argument and argument-argument priming.

In a following study, Hare, M. Jones, Thomson, Kelly, and McRae 2009 tested event knowledge activation even in more detail, by showing that nouns prime other nouns typically co-occurring as arguments in the same events. In particular, the authors used the SOA priming paradigm to demonstrate that:

— nouns of events prime participants (*sale-shopper*) and objects (*trip-luggage*) typically found at those events;

— locations prime people/animals and objects (*hospital-doctor,barn-hay*) typically found at those locations;

— instrument nouns prime things on which they are commonly used (*key-door*).

These effects on event-based priming (summarized in Figure 1.4) support the hypothesis of a mental lexicon arranged as *a web of mutual expectations*. Such expectations are encoded in the lexical representations, and the experimental subjects are clearly able to exploit them to compute the plausibility (i.e. the thematic fit) of the argument nouns as fillers of the roles of the verb (McRae, M. Spivey-Knowlton, and Tanenhaus 1998; Lenci 2011).[7]

The self-paced reading and the ERP experiments by Bicknell, Elman, Hare, McRae, and Kutas 2010 shed more light on how the mechanisms behind the generation and the dynamic update of the expectations during sentence processing, proving that they depend also on how other argument roles have been saturated. In particular, Bicknell and colleagues aimed at testing the hypothesis that the typicality of a filler of the patient role is a function of the filler of the agent role. For example, if the agent noun is *journalist*, *spelling* will be a likely filler for the patient role of the verb *to check*, while if the agent is *mechanic*, *engine* will be a more likely patient.

The authors showed how the expectations influence comprehension, by comparing the processing times between sentences with *congruent* arguments (e.g. *journalist* and *spelling*) and sentences with *incongruent* arguments (e.g. *mechanic* and *spelling*).

7. As pointed out by Lebani and Lenci 2017, the notion of thematic fit is closely related to the one of *selectional preferences*, which is probably still more popular in the fields of theoretical and computational linguistics. The principal difference between the two is that the former concerns the estimation of a continuous degree of compatibility between an argument and a thematic role, whereas the latter concerns the matching between the discrete semantic features of an argument and the feature(s) required by the predicate for the specific role.

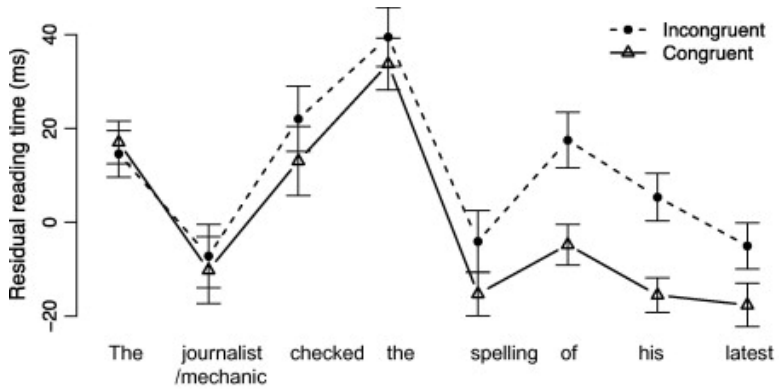


Figure 1.5. – Image taken from Bicknell, Elman, Hare, McRae, and Kutas 2010, showing the difference in the mean residual reading times between the congruent and the incongruent condition.

As a result, Bicknell and colleagues reported significantly shorter reading times for the congruent condition, as well as significantly smaller N400 amplitudes (see also Figure 1.5).

Moreover, results similar to those of Bicknell, Elman, Hare, McRae, and Kutas 2010 were later obtained in the self-paced reading and in the eye-tracking experiments by Matsuki, Chow, Hare, Elman, Scheepers, and McRae 2011, this time for congruent vs. incongruent instrument-patient pairs (e.g. *Donna used the shampoo to wash her filthy hair* vs. *Donna used the hose to wash her filthy hair*).

Finally, an ERP study by Metusalem, Kutas, Urbach, Hare, McRae, and Elman 2012 showed that the activation of the Generalized Event Knowledge is not limited to the linguistic items that would be licensed by the local linguistic context, reporting that even a contextually anomalous word elicited a reduced N400 component when it was generally related to a wider event scenario evoked by the discourse. For example, consider the case of the sentence *The scariest act is when the lion's tamer puts his head inside the lion's* ***clowns***, where the anomalous word in bold is related to

a circus scenario (it corresponds to the *event-related* condition in Figure 1.6).

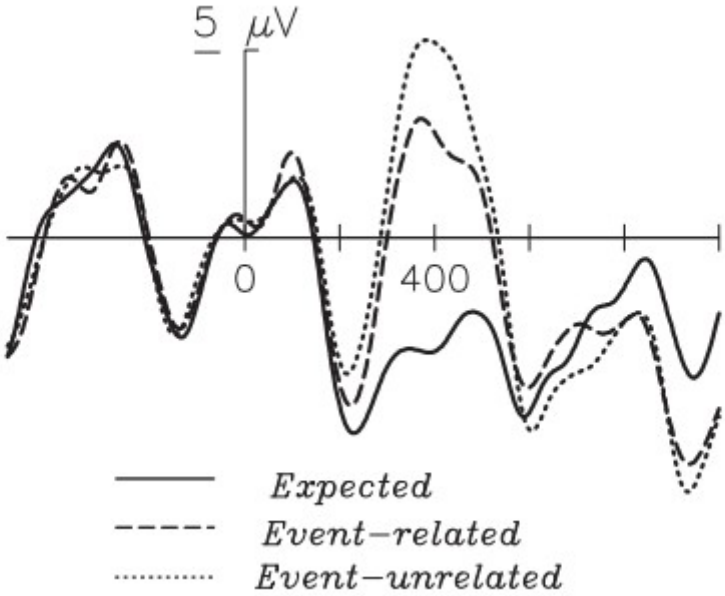


Figure 1.6. – Image taken from Metusalem, Kutas, Urbach, Hare, McRae, and Elman 2012, showing the grand average ERPs elicited by the target word at the midline parietal electrode. The three conditions in the experiment were: i) *expected* word; ii) word not licensed by the local context, but related to the event scenario evoked by the discourse (*event-related*); iii) word not licensed by the local context and not related to the event scenario evoked by the discourse (*event-unrelated*).

To sum up, the evidence presented in this section clearly points to the early activation of a Generalized Event Knowledge during sentence comprehension. The processing advantages due to arguments typicality cannot be attributed to a lexical-grammatical (i.e. purely linguistic) knowledge, as the experimental sentences did not include violations of combinatorial constraints of the lexical items (recall the Example 1c, in section 1.1), i.e. they do not describe *impossible* scenarios. Furthermore, such processing

advantages are still visible in anomalous sentences, in cases where the word causing the anomaly is related to the discourse context (Metusalem, Kutas, Urbach, Hare, McRae, and Elman 2012).

Differences like those reported between the congruent and the incongruent sentences of Bicknell, Elman, Hare, McRae, and Kutas 2010 and Matsuki, Chow, Hare, Elman, Scheepers, and McRae 2011 (as well as those between the sentences of the Example 1a-b) can be accounted for by our world knowledge. In other words, we are able to build semantic representations both for musicians playing in a theater and for gardeners playing in a cave, but only the first event type will be stored in our semantic memory as a plausible and typical scenario. On the other hand, sentences describing atypical situations will require more processing effort, since their semantic content presumably does not match any pre-stored representation.

1.3. Language Comprehension as a Cue-to-Category Mapping Problem

1.3.1. Words as Cues to Event Knowledge

The hypothesis that our semantic memory stores a Generalized Event Knowlegde (henceforth GEK: we adopt the acronym from McRae and Matsuki 2009) imposes a revision on traditional views on the comprehension process.

In such views, the comprehension of an input sentence is mainly seen as a building operation: it starts from the mapping from low-level features of the perceptual input to the sub-lexical and lexical level of representation;

then, lexical representations enter into a rule-driven compositional process at the syntactic-semantic level, leading to the construction of a more higher-level structure, corresponding to the meaning of the message (M. Brown and Kuperberg 2015). The lexicon, in such a perspective, has to be conceived as a sort of dictionary-like structure, i.e. a relatively stable inventory of entities with fixed meanings, which can be accessed by looking up specific entries (Elman 2014).

In the GEK view, unification-ready structures, i.e. the event representations, are assumed to be stored in the long-term memory and are activated by the lexical items during processing. Thus, the representation that is the target of the comprehension process does not necessarily need to be *composed*: if a representation for the event currently being processed is already stored in the lexicon, it just needs to be *retrieved*. As a consequence, the lexemes cannot be seen anymore as mere building blocks associated with a stable meaning, as their semantic content will inevitably depend on the context of the events that they contribute to activate (Rumelhart 1979; Elman 2009).

In his seminal work, inspired by the psycholinguistic findings of McRae and colleagues, Jeffrey Elman has synthesized these notions in the so-called *words-as-cues model of lexical knowledge* (Elman 2009; Elman 2014):

> *... suppose one views words not as elements in a data structure that must be retrieved from memory, but rather as stimuli that alter mental states (which arise from processing prior words) in lawful ways. In this view, words are not mental objects that reside in a mental lexicon. They are operators on mental states. From this perspective, words do not have meaning; they are rather cues*

to meaning. (Elman 2009).

Elman analyzed the findings of McRae and colleagues, arguing that such evidence could be accounted for only by a lexical knowledge that is extremely detailed, idiosyncratic and verb-specific, as the verb is the element providing the strongest cues to tap into our knowledge of events. What is striking, according to Elman, is not the fact itself that the expectations on an argument may change as a function of other arguments (e.g. the agent-verb combinations in the study by Bicknell et al., 2010), but rather the time course in which the update of the expectations is carried out. What is significant is that the effect occurs very early during processing, immediately at the argument following the verb: since it was generally assumed that immediate effects reflect lexical processing, while later ones reflect the late integration of semantic/pragmatic knowledge, the rapidity of the update would suggest that event information must be already embedded in the lexicon. After reviewing the results obtained by T. R. Ferretti, McRae, and Hatherell 2001, Matsuki, Chow, Hare, Elman, Scheepers, and McRae 2011 and Metusalem, Kutas, Urbach, Hare, McRae, and Elman 2012, Elman concludes that the influences of previously-saturated roles, of the verb aspect and of the discourse context all occur within the time frame that was supposed to be a prerogative of lexical information (Elman 2014). If one wanted to conceive the lexicon as a dictionary, this would lead to a combinatoric explosion of the entries that are necessary to account for the evidence presented above.

Consequently, Elman's theoretical proposal is a radical departure from the traditional view of the lexicon as an enumerative dictionary, and it consists of a **Simple Recurrent Network** (SRN: see the schematic illustration

in Figure 1.7) that is able to learn the connections between events and their typical participants, as well as typical sequences of action that unfold over time. What is particularly appealing about this model, according to Elman, is that the information flow is multidirectional and that there are feedback loops allowing the system to be affected by prior states. In such a model, instead of being conceived as a dictionary, lexical knowledge is thought to be encoded into the weights connecting the units within the network.

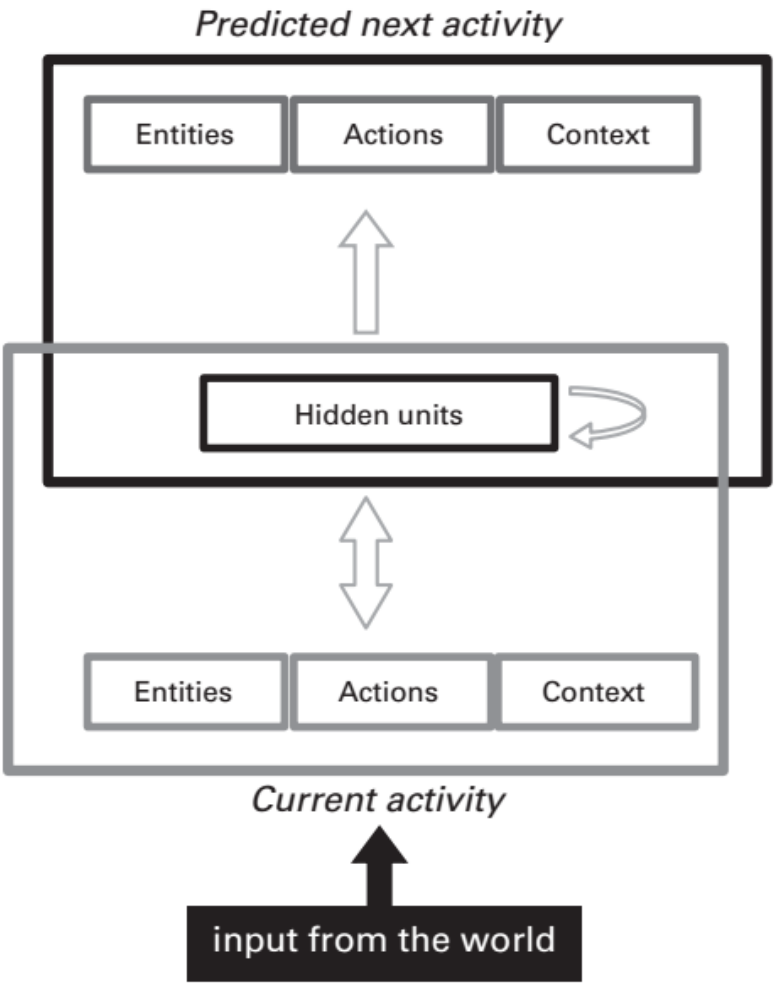


Figure 1.7. – Image taken from Elman 2014, illustrating how Elman's recurrent network model works. In the 'current activity' portion of the network, pattern of co-occurrence of typical events are learned (e.g. *The musician plays the flute in the theater*). The creation of recurrent connections between the entity-action-context units and the hidden units allows the network to predict the completion of input patterns (e.g., *musician* and *play* activate *flute* and *theater*). The 'predicted next activity' portion learns typical sequences of actions, and allows to anticipate subsequent activities.

The event knowledge activated during sentence processing is at the core of Elman's framework:

Events plays a central role in organizing our experience. Event

> *knowledge is used to drive inference and access memory, and it affects the category we construct. An event may be defined as a set of participants, activities and outcomes that are bound together by causal interrelatedness.* (Elman 2009)

In other words, lexical items are seen as cues to event knowledge, and the 'meaning' consists precisely of the events that they activate. Starting from linguistic input, the goal of the model is to learn to identify the mutual constraints between entities and actions. Notice that events are not just temporary activities (e.g. *The man cuts the meat in a restaurant with a knife*), but also more structured sequences of actions (e.g. *The man goes to the restaurant, he orders some meat, he cuts it with a knife, he eats it and he finally asks for the bill* etc.).[8] During learning, the model is likely to see only a subset of the actions/participants that make up events, but over time it will manage to learn generalizations between and across events, and to fill in missing information when necessary.

Words are not the only cues allowing access to event knowledge: abstract grammatical constructions, the discourse context and various types of non-linguistic information contribute as well in the activation process. Finally, as the new information comes in, independently from its nature (syntactic, semantic, pragmatic etc.), the activation pattern consequently changes. Indeed, Elman's SRN model is a constraint-based one: all available information is immediately integrated, and its processing can be seen as adding a new set of constraints on the possible interpretations of the sentence.

8. From this point of view, a similar notion is that of *script*, which has been widely used in Artificial Intelligence and Cognitive Science (Schank and Abelson 2013).

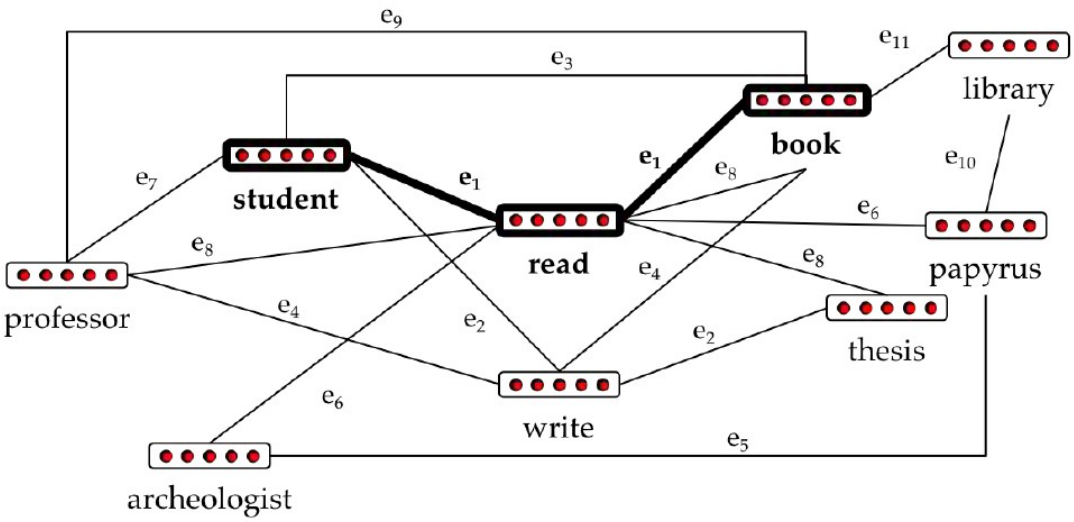


Figure 1.8. – Representation of the GEK as a network of events and participants. Edges between words are assumed to be co-activation links.

Elman's proposal bears some resemblance with the **situation models** of discourse comprehension, firstly introduced in the experimental psychology literature in the 80s (Van Dijk and Kintsch 1983; Kintsch 1988; Zwaan and Radvansky 1998; Zwaan and Madden 2005): perhaps the most striking point of similarity lies in the relation between the linguistic items and the operation of retrieval of previous experiences from the semantic memory. More explicitly, all these theories have in common their framing of sentence comprehension as a *cues-to-category mapping problem* (Qian, Jaeger, and Aslin 2012).

Situation models generally assume that language is a set of instructions used to create a mental representation of a context that is currently being described, and the storage and retrieval of previously experienced situations plays an essential role in such a process. An example of a proposal inspired by such principles is the exemplar-based model of sentence pro-

cessing by Johns and M. N. Jones 2015. According to this model, sentences occurring in natural language are stored as *memory traces*, and processing is seen as a memory task where the retrieval is driven by the similarity between a probe string and the exemplars. A dynamic data structure provides a representation of the current context, which gets updated when new words are processed. The more an exemplar representation of a sentence is similar to the content of a such a structure, the more it will be activated, and the most strongly activated exemplars will be used, in turn, to generate expectations on the upcoming input.

The authors also presented a computational implementation by means of a Random Permutation vector space model (Sahlgren, Holst, and Kanerva 2008), managing to capture a wide range of behavioral effects (reduced relative clause processing, event knowledge activation etc.). Interestingly, such results were obtained on the basis of just storage and retrieval of vector encodings of sentences occurring in written text, with no abstract representation computed over the stored exemplar, thus suggesting that the above-mentioned operations could account for a surprising amount of contextual variability.

1.3.2. Hierarchical Generative Frameworks of Language Processing

A similar re-framing of the comprehension problem has been recently proposed by language processing studies on *hierarchical generative frameworks* (see Kuperberg and Jaeger 2016 for an overview). In the same way as the GEK-related hypotheses, such models do not assume that comprehensions consists of assembling higher-order representations from lower-level building blocks.

Generative frameworks, instead, take as their starting point the fact that a comprehender tries to infer the underlying *cause* of an observed set of sensory inputs that unfolds over time (M. Brown and Kuperberg 2015; Kuperberg 2016). The process is top-down, in the sense that, in order to carry out the task, the comprehender relies on a *generative model*, a set of representations hierarchically-organized in levels (they correspond to the linguistic units of analysis: phonemes, lexemes, syntactic and semantic structures, the discourse context etc.). At any given time during processing, and given the current hypotheses (*beliefs*) of the comprehender, the generative model is taken as the best explanation of the observed input.

The higher-level hypotheses are tested by generating probabilistic predictions that are propagated to lower levels, consequently changing the distribution of prior beliefs at these levels before new input becomes available. As soon as new information is processed, the comprehender learns whether the top-down predictions are supported or not. In the context of sentence comprehension, for example, predictions correspond to expectations on the upcoming words, which, in turn, can lead to expectations on their acoustic and phonetic features (see also the generative framework

for speech processing by Kleinschmidt and Jaeger 2015). New bottom-up evidence that is consistent with the predictions (e.g. a predictable word) is immediately 'explained away', while the discrepancies between evidence and predictions (*prediction error*) are propagated back up the generative model, allowing to update the higher-level hypotheses about the underlying message.

By means of an iterative cycle between probabilistic inferences and predictions, the comprehender manages to gradually reduce the prediction error and converge on the higher-level representation that causes and best explains the full set of observations. Similarly to the words-as-cues framework, comprehension is conceived as a form of cue-to-category mapping, as the goal is to infer the message that the speaker intends to communicate, that is, a higher-level category explaining the linguistic cues in the input.

It should be pointed out, however, that such a category does not correspond necessarily to a specific event scenario stored in the semantic memory: as stated by Kleinschmidt and Jaeger 2015 an 'ideal adapter' should also be able of 'generalizing to the similar'. Some of the observed context may not provide enough evidence to support the retrieval of a specific event, but they may provide enough evidence to infer a more general *event structure*, a representation that would be still higher on the generative hierarchy. For example, we may hear a sentence where the participants filling the roles are novel with respect to our past experiences with the same event type, i.e. it does not match any specific event representation in our semantic memory. However, we might have inferred an abstract event structure from similar events, in which we have just the specifications of the coarse-grained semantic properties that are necessary for a participant

to match the role requirements. In such cases, adaptation would be carried out by checking the semantic features of the new role fillers. [9]

Event structures allow comprehenders to formulate strong predictions on the arguments of a sentence, e.g predictions on the animacy of an upcoming argument can be generated on the basis of the selectional restrictions of verbs (Paczynski and Kuperberg 2011; Paczynski and Kuperberg 2012), or on the combination of the animacy and the linear order of the other arguments (Weckerly and Kutas 1999; Paczynski and Kuperberg 2011). In all these cases, N400 amplitude has been shown to be modulated by the fulfillment of the prediction.

1.4. Conclusion

In this first chapter, we proposed a review of the most recent advances in the sentence processing literature, particularly concerning the role of event knowledge in the comprehension process.

First, we have presented the issue of the balance between storage and computation as primary in the theoretical challenge of explaining natural language productivity (Baggio and Hagoort 2011). Evidence coming from studies on Event-Related Potentials (ERP) -and in particular on the N400 component- suggest that there are processing differences also between fully understandable sentences, depending on the *typicality* of the situation that they describe. While the focus of linguistic theory was previously

9. Recent psycholinguistic and neurolinguistic studies have advanced the hypothesis that the *selectional restrictions* of the verbs are actually coarse-grained abstractions built out of exemplar-type event representations in the semantic memory, such as Agent $< animate >$ - Action Verb - Patient $< inanimate >$ (Paczynski and Kuperberg 2012; Warren, Milburn, Patson, and Dickey 2015). In the same way as prototypical representations of categories are computed by abstracting away from accumulated experiences with category members (Medin, Altom, and Murphy 1984), such coarse-grained abstractions would exist alongside the more specific event memories.

on rule-driven compositional operation as the only possible answer to the productivity issue, such findings have caused a shift of interest towards the unification-ready structures that are stored in the semantic memory.

In the second section, we have reviewed a series of experimental studies presenting strong evidence for the activation of a Generalized Event Knowledge during sentence processing (McRae and Matsuki 2009). Contrary to the claims of traditional modular theories, which assumed the primacy of 'linguistic-only' information in the comprehension process, such studies prove that humans *immediately* activate their knowledge about events and typical participants and use it to anticipate the upcoming input. Sentences describing situations that are consistent with our event knowledge have been shown to be easier to process (Bicknell, Elman, Hare, McRae, and Kutas 2010; Matsuki, Chow, Hare, Elman, Scheepers, and McRae 2011).

The final section of the chapter has been dedicated to the presentation of some recent proposals that try to re-frame the problem of sentence comprehension, which have in common a change of perspective on the process of building a representation for the meaning of the sentence:

— in the *Words-as-Cues Framework* (Elman 2009; Elman 2014), words are seen as stimuli that alter mental states. Rather than units carrying a stable meaning, they are cues activating events in the GEK. As the processing of an input sequence unfolds, words gradually constrain the interpretation of the sentence, that is, the event or the situation the sentence refers to;

— the *Hierarchical Generative Models* (Kleinschmidt and Jaeger 2015; Kuperberg and Jaeger 2016; Kuperberg 2016) assume that the goal of language processing is finding the latent cause of an observed set of

stimuli. This 'cause' is a higher-level category (the *generative model*), corresponding to the message that the speaker intended to convey. The comprehender has to retrieve it by making probabilistic predictions on the upcoming input, by checking their consistency as the new input comes in and eventually updating the beliefs on the generative model. This iterative cycle ultimately ends when the comprehender has retrieved the higher-level category that best accounts for the stimuli in the input.

Some of the concepts that we have introduced in this chapter will be also at the core of our model. More specifically, we propose that:

1. human semantic memory stores knowledge about events and their typical participants. In other words, we assume that *the* GEK *is the content of the Memory component, and that this content has not to be considered language-specific, but it is acquired from both first-hand, sensory-motor experience of events and linguistic experience*;
2. events in the GEK are dynamically activated during sentence processing and used to anticipate the upcoming input. As soon as new input becomes available, we update our predictions on the event or situation that the current message is likely to communicate;
3. consistency between stored knowledge and the actual input is one of the parameters determining processing complexity/difficulty. Since processing differences have been shown for sentences differing for the fitness of an argument, *we consider thematic fit as an inverse index of the cost of semantic unification*. For sentences containing atypical argument combinations, it is more difficult to build a semantic representation of the situation described by the sentence.

In the following chapter, we are going to introduce the general computational framework which we used for computational modeling, i.e. Distributional Semantic Models (DSMs). A specific attention will be dedicated to DSMs making use of syntactic structure and to their recent applications to psycholinguistics modeling.

2. Methodology

Main contributions of the chapter:

— we offer a general overview on Distributional Semantic Models: historical evolution, main principles, typologies;

— we discuss the applications of Distributional Semantics to psycholinguistics and cognitive modeling, with a specific focus on thematic fit models;

— we present the main vector-based approaches to the compositionality problem.

2.1. Vector Space Models for Semantics

2.1.1. From the Distributional Hypothesis to Word Vectors

Distributional Semantic Models (DSMs) have by now become a standard *de facto* in the Natural Language Processing community, as they provide an efficient and data-driven methodology for building semantic representations. All the DSMs rely on some version of the so-called *Distributional Hypothesis* (Lenci 2008; Sahlgren 2008), which can be stated as follows:

> *The semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B occur.*

The roots of such hypothesis are grounded in the tradition of studies of the American structuralism, and in particular in the work by Zellig Har-

ris. Similarly to other structuralists, like Bloomfield, Harris was skeptical about the possibility for the linguistic enterprise to directly investigate meaning in all its social manifestations. However, although linguistic facts are considered to be influenced by extralinguistic factors that are beyond the reach of a linguistic theory, the distributions of those facts can still be object of investigation.

In *Mathematical Structures of Language* (Harris 1968), Harris explained that, insofar as meaning has a dimension that is internal to language, it must be susceptible to distributional analysis. Moreover, in *Distributional Structure* (Harris 1970), he talked about the existence between similarities/dissimilarities in meaning and similarities/dissimilarities in the related linguistic distributions. In other words, independently of the nature of meaning, it is the *distributional similarity* of the linguistic items that should be at the center of the linguist's interest (see also Sahlgren 2008). Thus, nothing can be said about the meaning itself: the focus of such an approach is on how much similar/dissimilar the meanings of words are.

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

Intuitively, the distributional representation of a lexeme coincides with its co-occurrence frequencies with a given list of contexts, which could be, in the simplest case, other lexemes. Suppose that we want to represent the words *dog* and *cat* and that our corpus is composed just by the sentences

below:

— My house is full of cats and dogs.

— The dog bit a man on the street.

— Tonight, my dog was barking all the time.

— This cat can eat and purr at the same time.

— Your dog should stop eating everything.

— Everybody in this street has a cat.

— I'd like to spend some time alone in my house, with my dogs and my cats.

As contexts, we select all the nouns and the verbs occurring in these sentences (including the target words themselves), and we represent the two words as vectors as in Table 2.1, inserting in each column a value corresponding to the co-occurrence of the target word with a specific lexeme. In this basic example, a single co-occurrence is counted when two words co-occur within the same sentence.

Target word	house	time	street	bite	stop	eat	purr	bark	like	dog	cat
dog	2	2	1	1	1	1	0	1	1	0	2
cat	2	1	1	0	0	1	1	0	1	2	0

Table 2.1. – Distributional Space - Example

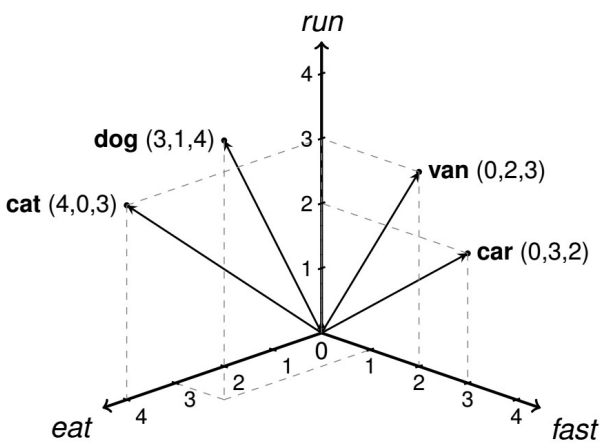


Figure 2.1. – Representation of a distributional space with three contextual dimensions. The image is taken from Lenci, 2014 (oral presentation at the Global Wordnet Conference, Tartu).

Normally, the semantic similarity between lexemes is assessed by means of some *vector similarity metric*. The most popular metric in the literature is by far the *vector cosine* (Turney and Pantel 2010), which consists of the cosine of the angle between the word vectors, ranging from 0 (the words are maximally dissimilar, as they have no contexts in common) to 1 (the two words have exactly the same meaning) [1].

The cosine similarity between two word vectors is computed as follows:

$$sim(\vec{A}, \vec{B}) = cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \tag{2.1}$$

which is equivalent to

$$sim(\vec{A}, \vec{B}) = cos(\theta) = \frac{\sum_{k=1}^{n} A(k)B(k)}{\sqrt{\sum_{k=1}^{n} A(k)^2}\sqrt{\sum_{k=1}^{n} B(k)^2}} \tag{2.2}$$

1. Cosine similarity is typically used in positive spaces, where the values are bounded in the $[0, 1]$ interval.

where $\vec{A}$ and $\vec{B}$ are the word vectors being compared, while n is the vector length. In the example in Table 2.1, the similarity between the *dog* and the *cat* vectors would be: $sim(\vec{dog}, \vec{cat}) = 0.411652$.

The cosine has the advantage that length differences between the vectors are irrelevant, since only their angle matters: this is an important property, as similar words might greatly differ in frequencies, and consequently their vector would have very different lengths. Proposals for alternative measures have been advanced in the literature, such as, for example, metrics based on the weighted overlap between the features of the vectors (Pilehvar, Jurgens, and Navigli 2013; Camacho-Collados, Pilehvar, and Navigli 2015; Santus, Chiu, Lu, Lenci, and C.-R. Huang 2016; Santus, Chersoni, Lenci, and Blache 2017; Santus, H. Wang, Chersoni, and Y. Zhang 2018), but overall vector cosine is still the most commonly chosen metric.

2.1.2. The First Generation of DSMs: The Count Models

There are two principal ways in which DSMs can be constructed.

The first and the most traditional methodology consists of initializing the word vectors with co-occurrence counts (therefore, models built in this way are also known as *count models*, according to the terminology introduced by Baroni, Dinu, and Kruszewski 2014), as we did in our toy example in Table 2.1. Informally speaking, in such a procedure it is necessary to identify the following elements:

1. the **target** linguistic elements that need to be represented;
2. the **linguistic contexts** that will constitute to the vector dimensions;
3. what counts as a **co-occurrence event**; in the most common case,

this is the co-occurrence of the target and the context in the same sentence, or within a word window of fixed size (but see also Section 2.1.3);

4. eventually, the **weighting schemes** and the **transformations** that are applied to the semantic space.

Point 4 is optional, but it is still a common practice to apply some transformations to raw co-occurrence frequencies, as this has been shown to greatly improve the performances of DSMs (Turney and Pantel 2010; Bullinaria and J. Levy 2012; O. Levy, Y. Goldberg, and Dagan 2015).

An extremely common weighting scheme in the literature is the **Pointwise Mutual Information** (PMI) (Church and Hanks 1990; Pantel and Lin 2002), which is used to reduce the bias to higher frequency contexts and to assign a higher value to the contextual dimensions having a strong statistical association with the target.

Given a target t and a context c, and their respective probabilities of occurrence in a training corpus $P(t)$ and $P(c)$, we will define their PMI as:

$$PMI(t, c) = \log \frac{P(t, c)}{P(t) * P(c)} \tag{2.3}$$

$P(t, c)$ is the joint probability of t and c that is observed in the training corpus. In case t and c are statistically independent, we expect this probability to be equal to the product of the probabilities of the single events. Dimensions corresponding to the co-occurrences of statistically independent events will be zeroed by the PMI, as the co-occurrence is due to random chance and to the existence of a statistical association between target and context. On the other hand, higher values will be assigned to those contexts whose co-occurrence with the target is above random chance (i.e.

$P(t,c) > P(t) * P(c))$ [2].

A well-known problem with PMI is that it is biased towards rare events. An alternative weighting scheme that partially corrects the problem is the Local Mutual Information (LMI; Evert 2004), in which the PMI association score of two events is simply multiplied by their joint frequency:

$$LMI(t,c) = f(t,c) * PMI(t,c) \tag{2.5}$$

LMI has been successful especially in studies dedicated to building DSMs enriched with dependency information and it is the weighting scheme adopted by some of very popular *structured* DSMs (see Section 2.1.4), such as Distributional Memory (DM; Baroni and Lenci 2010).

Another operation that is often applied to the semantic space is some type of **dimensionality reduction**, in order to limit the number of the vector components. Indeed, many of the contexts have very little discriminative power, and thus several linear algebra-based techniques have been proposed to smooth the target-context matrix.

The most popular one is **truncated Singular Value Decomposition** (SVD) (Deerwester, Dumais, Furnas, Landauer, and Harshman 1990; Landauer and Dumais 1997), also called Latent Semantic Analysis (LSA) when it is applied to word similarity (Turney and Pantel 2010).

SVD is a matrix decomposition technique that creates a low-dimensional linear mapping between the targets (the rows of the space) and the con-

2. It is a common pratice to use the **Positive Pointwise Mutual Information** (PPMI), in order to keep non-zero values only for the dimensions that are positively associated with the target. It is computed as:

$$PPMI(t,c) = max(PMI(t,c), 0) \tag{2.4}$$

texts (the columns). Given a target-context matrix M or rank d, M is first decomposed into the product of three matrices $U\Sigma V^T$, where U and V are orthogonal matrices and Σ is a diagonal matrix. Then, let Σ_k be the diagonal matrix formed by the top k singular values in Σ ($k < d$), and let U_k and V_k be the matrices obtained by selecting the corresponding columns from U and V: $U_k\Sigma_k V_k^T$ is the matrix of rank d that better approximates the original matrix M. The projection into a lower-dimensional space has the effect of discovering a number of 'latent' (hidden) dimensions, the ones which capture the bigger amount of variance in the original data, while reducing noise.

For practical purposes, PMI (or PPMI) and SVD are often applied together in count-based DSMs, and several papers have shown that this combination is particularly effective in many similarity-related tasks (Bullinaria and J. Levy 2012; Santus, Chersoni, Lenci, C.-R. Huang, and Blache 2016). Furthermore, refinements of the original SVD have been proposed in the literature, leading to better results on word similarity (O. Levy, Y. Goldberg, and Dagan 2015; Sahlgren and Lenci 2016).

2.1.3. The New Generation: The Predict(ive) Models

The recent years have seen the introduction of a new family of DSMs, which has been developed within the deep learning research community. This new generation of word vectors formulate the problem of assigning a vector representation to words as a supervised task: the values of the vector dimensions are set in order to maximize the probability of the context co-occurring with the target words in the corpus (Baroni, Dinu, and Kruszewski 2014). Such models are also known as *predict(ive)-models*, as

the vector weights, instead of being initialized with co-occurrence counts, are set in order to optimize the prediction of co-occurring contexts.

The vectors obtained by means of supervised training are alternatively called *word embeddings* or *neural embeddings*, and have the advantage of enabling efficient computation of word similarities in low-dimensional spaces (O. Levy and Y. Goldberg 2014a). Moreover, the procedure for building predict-based representations is more compact (instead of several steps of frequency collection, application of weighting schemes and/or of dimensionality reduction techniques etc., training initialization is all that is needed) and they have been reported to outperform the traditional count-based models in a wide variety of tasks (Baroni, Dinu, and Kruszewski 2014; but see also the discussion sections in O. Levy, Y. Goldberg, and Dagan 2015 and in Sahlgren and Lenci 2016).

Among the state-of-the-art embedding methods, the most famous ones are probably the **Continuous Bag of Words** (CBOW) and the **Skip Gram**, introduced by Mikolov and colleagues and implemented in the Word2Vec software (Mikolov, Chen, G. Corrado, and Jeffrey Dean 2013; Mikolov, Sutskever, Chen, G. S. Corrado, and Jeff Dean 2013).

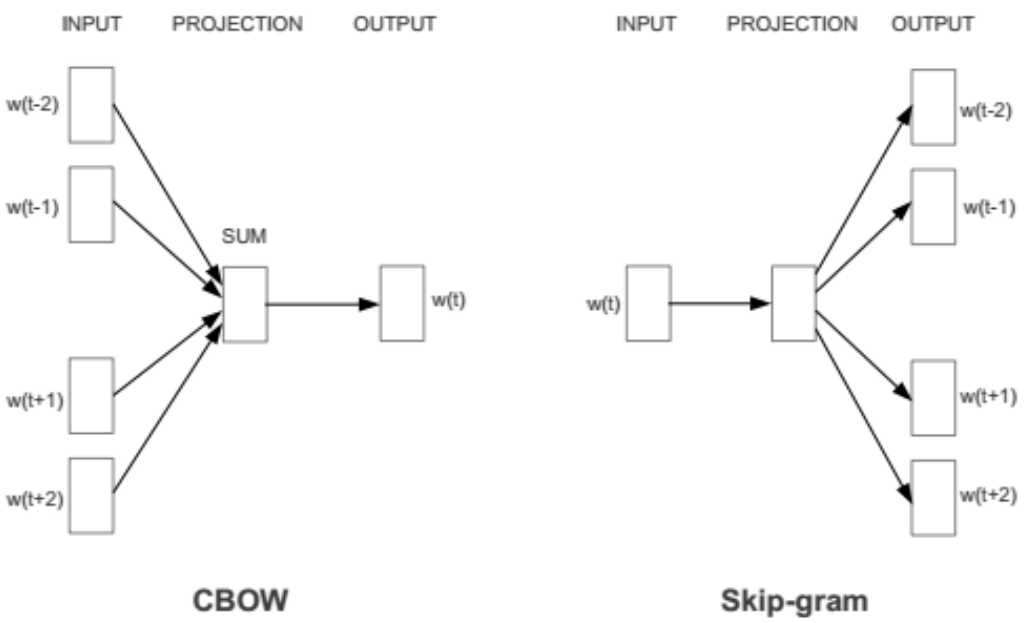


Figure 2.2. – Scheme of the network architectures of the Continuous Bag-of-Words (CBOW, on the left) and of the Skip Gram (on the right). Image taken from Mikolov, Chen, G. Corrado, and Jeffrey Dean 2013.

In both the Word2Vec architectures, the values of the word vectors are latent and treated as parameters to be learned. The neural networks of the two models have a rather simple structure, with a single hidden layer.

In the CBOW architecture, the model has to predict the word in the middle of a symmetric window, on the basis of the sum of the vector representations of the words in the window, whereas in the Skip-gram architecture the model has to predict the surrounding context words given a word occurring at the center of the window.

In the original studies by Mikolov and colleagues, the training of word vectors was carried out by minimizing a cost function by means of the *gradient descent* algorithm [3], which is widely known in neural network re-

3. This is not always true for all the training functions available in the Word2Vec package: for

search.

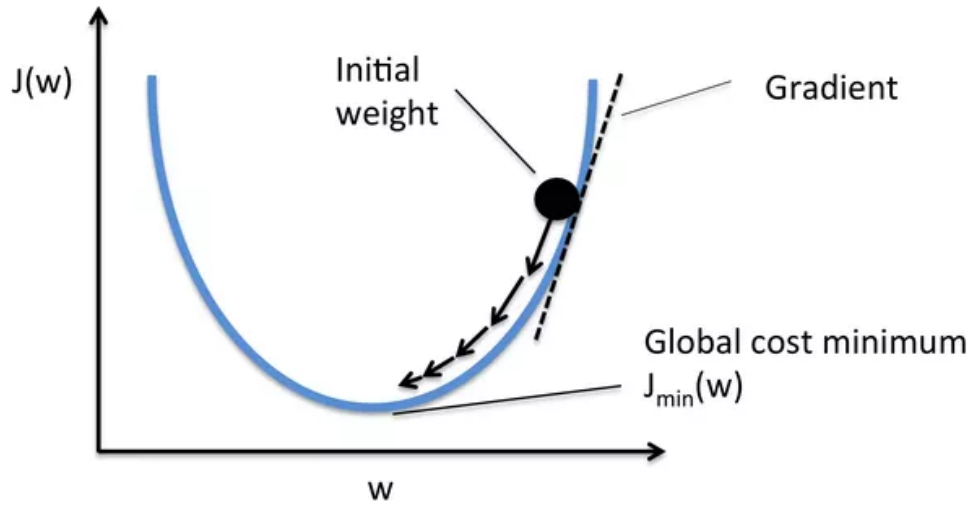


Figure 2.3. – Illustration of gradient descent with a single parameter w to be trained on the x-axis. On the y-axis, the cost function$J(w)$ to be minimized. Starting from the initial weight, the weight w is updated at each iteration until the minimum value for $J(w)$ has been reached. The gradient descent algorithm stops when the parameter values do not change anymore, i.e. the global minimum of the function has been reached.
The image is taken from Sebastian Raschka's Deep Learning Blog. [4]

A subsequent study by Levy and Goldberg (O. Levy and Y. Goldberg 2014b) showed that the popular training method of the Skip-Gram with negative sampling could be casted as matrix factorization, and that the objective function of the network was implicitly factorizing a word-context PMI matrix, shifted by a constant offset. This was a particularly interesting achievement, because it provided a strong connection link between the two study traditions in DSMs, as well as new insights on how to improve the performances of count-based models on the basis of the practices used

example, the Negative Sampling is a reward function (the objective is to maximize the dot product of word-context vector pairs that have been observed in the corpus data and to minimize the dot product for randomly-generated pairs) and the optimizing algorithm is actually a *gradient ascent*.

4. https://sebastianraschka.com/faq/docs/closed-form-vs-gd.html.

in neural network training (O. Levy, Y. Goldberg, and Dagan 2015).

Currently, the predict-based models have estabilished themselves as a new standard in NLP. The performance advantage reported by Baroni, Dinu, and Kruszewski 2014 has been shown to be due to the optimization of some hyperparameters that had been already carried out by the system designers of the Word2Vec package, whereas in the case of count-based models the tuning has to be done manually (O. Levy, Y. Goldberg, and Dagan 2015). Nonetheless, the speed and the ease of training of Word2Vec models, with respect to the multiple processing steps that are needed to setup the count-based ones, are the factors that contributed the most to their success among researchers in Distributional Semantics. Starting from the works by Mikolov and colleagues, other standard architectures for training word embeddings have been proposed in the literature. Among these, the most popular ones are probably GloVe (Global Vectors; Pennington, Socher, and C. Manning 2014), based on a weighted least squares model trained on a global matrix of word-word co-occurrence counts, and the recently-introduced FastText (Bojanowski, Grave, Joulin, and Mikolov 2016), which exploits the idea of learning word representations as the sum of the embeddings of the character n-grams composing them, and thus it is more efficient than Word2Vec in modeling syntactic analogies.

However, it should be pointed out that predict-based models are not always the best choice: as it has been shown by previous studies, word embeddings generally have difficulties when the size of the training corpus is relatively small (Asr, Willits, and M. Jones 2016; Sahlgren and Lenci 2016; Asr and M. Jones 2017), or in tasks involving more structured linguistic knowledge (such as the *thematic fit estimation* task: see Baroni, Dinu, and

Kruszewski 2014 and A. Sayeed, Greenberg, and Demberg 2016).

2.1.4. The Choice of the Context

Until now, the examples of both count-based and predict-based models were limited to the so-called *bag-of-words* or *unstructured* DSMs, i.e. models conceiving the context as a (symmetric) *word window* surrounding the target word. Generally, a fixed value for the window width is selected and all the words occurring within the window contribute to the contextual representation of the target word, independently of the existence of a structural relation with it. [5]

But this **unstructured** notion of context is not the only possible one. Indeed, there also DSMs relying on *structured* contexts (the so-called **structured DSMs**), where the statistics of co-occurrence between words are collected in the form of corpus-derived triples: the word pairs and the lexical pattern or the relation linking them (Baroni and Lenci 2010). The relations can be either the syntactic dependencies extracted by a parser (G. Grefenstette 1994; Lin 1998; S. Padó and Lapata 2007; Baroni and Lenci 2010) or the roles extracted by a semantic role labeler (A. Sayeed and Demberg 2014; A. Sayeed, Demberg, and Shkadzko 2015) (see the Examples in Table 2.2 and 2.3).

Target word	Relation	Context Word	Frequency
dog-n	direct object	see-v	2
cat-n	subject	purr-v	8
professor-n	indirect object	send-v	4
good-j	adjectival modifier	friend-n	3

Table 2.2. – Distributional Space based on dependencies

5. Typically, grammatical words are discarded in such a process, and only co-occurring lexical words are retained as they are more informative of the meanings of the lexemes.

Target word	Relation	Context Word	Frequency
knife-n	instrument	cut-v	5
player-n	agent	score-v	2
pizza-n	patient	eat-v	9
park-n	location	skate-v	1

Table 2.3. – Distributional Space based on semantic roles

In the same way as traditional DSMs, weighting schemes and transformations can be applied to the semantic space.

Structured DSMs lead to more sparse spaces (i.e. if only the words in a syntactic/semantic relation are considered as contexts, the natural consequence is that fewer vector entries will have non-zero values), but on the other hand just the words with an actual relation with the target are included in its semantic representation.

In the same way as count-based models, also predict-based models can exploit structured contexts: O. Levy and Y. Goldberg 2014a showed how the training on bag-of-words contexts of the traditional Skip-Gram architecture can be carried out on arbitrary contexts, such as words in a specific dependency relation, independently from their linear distance in the sentence (e.g. in the sentence *Australian scientist discovers star with telescope*, the contexts for the word *scientist* would be *australian-adjectival modifier* and *discovers-*$subject^{-1}$). [6]

Compared to their unstructured relatives, structured DSMs have been shown to model a different kind of similarity (Turney 2012; O. Levy and Y. Goldberg 2014a; Hill, Reichart, and Korhonen 2016):

— unstructured DSMs model *domain similarity* (also known as *semantic relatedness*), i.e. the similarity between words that tend to occur to-

6. Notice that, in the typical notation for Structured DSMs, rel^{-1} denotes the inverse relation of rel.

gether as they are both related to the same topic, although they do not refer necessarily to similar concepts (e.g. *coffee* and *cup*);

— structured DSMs model *function similarity*, i.e. the similarity between words that tend to have similar functional relations to other words (e.g. *pizza* and *pasta*, as they are both typical patients of *to eat, to cook* etc.).

Therefore, structured DSMs are usually the first choice for tasks related to word categorization (Lapesa and Evert 2017), property generation, modeling of selectional preferences etc. (Baroni and Lenci 2010).

In our framework, we are particularly interested in modeling event knowledge, a problem that is connected to the selectional preference one, since it implies the estimation of the typicality of the arguments of a verb as event participants. In our approach, we will use syntactic dependencies as a surface approximation of the relations between event participants.

In the following section, we will review the studies in the computational psycholinguistics literature that make use of DSMs, with a specific focus on those aiming at modeling the evidence for the Generalized Event Knowledge (see Section 1.2), mostly on the basis of structured contexts. Also related to the topic of sentence understanding, we will look at how research on Distributional Semantics has tackled the problem of sentence representation.

2.2. Distributional Semantics and Cognitive Modeling

The Distributional Hypothesis can be interpreted in different ways. It can be seen, as in the traditional view by Harris, as a working assumption to investigate meaning, without making any further claim about the relation between semantic representations and distributional patterns. Since the early 90s, an alternative and more radical view on such a relation came from the research in cognitive science, when Miller and Charles 1991 proposed the so-called 'strong version' of the Distributional Hypothesis (Lenci 2008). According to these authors, repeated encounters with words in different contexts lead to the formation of an abstract representation of the word meaning, derived from the most significant contexts of usage. From this point of view, the Distributional Hypothesis become a cognitive hypothesis, as the distributional behavior of words determines their semantic content at the cognitive level.

Indeed, some of the most influential distributional models have been developed as cognitively plausible models of language learning (LSA: Landauer and Dumais 1997; HAL: Burgess and Lund 1997), and since their earlier days they have often been used to model psychological phenomena such as similarity judgements and semantic/associative priming (Rubenstein and Goodenough 1965; Miller and Charles 1991; M. N. Jones, Kintsch, and Mewhort 2006; Ettinger and Linzen 2016; Lebani and Lenci 2017).

Therefore, it is not a surprise that DSMs have seen an increasing popularity also in psycholinguistic modeling: beyond the ease in obtaining distributional representations of linguistic units by using the many available software libraries, the results from the experimental studies constitute

an appealing benchmark for the modeling ability of the DSMs. Indeed, some recent proposals for the evaluation of such models aimed at modeling data collected through different experimental paradigms, such as eye-tracking (Søgaard 2016) and event-related potentials (ERP: Ettinger, Feldman, Resnik, and Phillips 2016).

2.2.1. Distributional Models for Event Knowledge and Thematic Fit Modeling

Distributional modeling of event knowledge has always had a very strong relationship on the studies on thematic fit, as the latter provided the gold standard datasets on which the former have to be tested. Moreover, for compositional distributional models, the notion of thematic fit is easier to model with vector spaces than the similar notion of 'selectional preference', which is perhaps more common in theoretical linguistics. Indeed, selectional preferences concern the matching between the discrete feature requirements for a verb-specific role and a given argument, while thematic fit can be described as a 'degree of compatibility' between the two. And, whilst distributional similarity is a good proxy for the latter notion, vector representations notoriously have difficulties in handling discrete knowledge (Smolensky 1990; Fodor and Lepore 1999; Gupta, Boleda, Baroni, and S. Padó 2015; Gupta, Boleda, and S. Padó 2017).

Most research work on thematic fit estimation has focused on *count-based* vector representations since, in their comparison between such models and the low-dimensional neural embeddings, Baroni, Dinu, and Kruszewski 2014 found that thematic fit estimation is 'the only benchmark on which predict models are lagging behind state-of-the-art performance'. This is

also in line with A. Sayeed, Greenberg, and Demberg 2016's hypothesis that 'thematic fit modeling is particularly sensitive to linguistic detail and interpretability of the vector space'.

Erk, S. Padó, and U. Padó 2010 were, at the best of our knowledge, the first authors to measure the correlation between human-elicited thematic fit ratings and the scores assigned by a syntax-based DSM. More specifically, their gold standard consisted of the human judgements collected by McRae, M. Spivey-Knowlton, and Tanenhaus 1998 and U. Padó 2007. The plausibility of each verb-filler 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 evaluated their Distributional Memory (DM) framework on the same datasets, adopting an approach to the task that has become dominant in the computational linguistics literature: for each verb role, they built a prototype vector by averaging the dependency-based vectors of its most typical fillers (see also Figure 2.4). The higher the similarity of a noun with a role prototype, the higher its plausibility as a filler for that role. Lenci 2011 has later extended the 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 Bicknell dataset (Bicknell, Elman, Hare, McRae, and Kutas 2010).

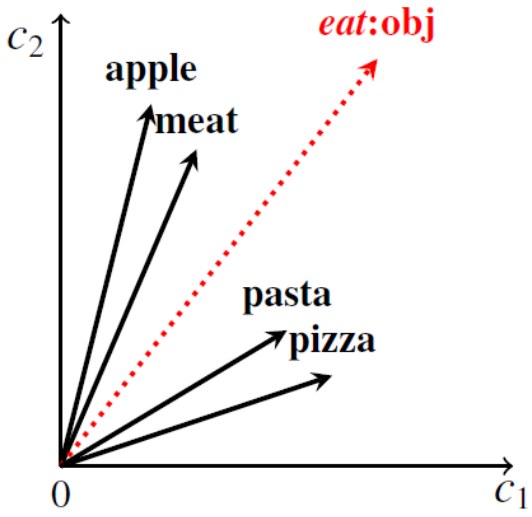


Figure 2.4. – Example of a prototype vector for the direct object of *to eat*, computed by averaging the typical fillers.

The thematic fit models proposed by A. Sayeed and Demberg 2014 and A. Sayeed, Demberg, and Shkadzko 2015 are similar to Baroni and Lenci's, but their DSMs were built by using the roles assigned by the SENNA semantic role labeler (Collobert, Weston, Bottou, Karlen, Kavukcuoglu, and Kuksa 2011) to define the feature space. These authors argued that the prototype-based method with dependencies works very well when applied to the agent and to the patient role, which are almost always syntactically realized as subjects and objects. However, they claim, it might be problematic to apply it for estimating the thematic fit for different roles (such as instruments and locations), since the construction of the prototype would have to rely on prepositional complements as typical fillers, and the meaning of prepositions can be ambiguous. Comparing their results with Baroni and Lenci 2010, the authors showed that their system outperforms the syntax-based model DepDM and almost matches the scores of the best

performing TypeDM, which uses hand-crafted rules. Moreover, they were the first to evaluate thematic role plausibility for roles other than agent and patient, as they computed the scores also for the instruments and for the locations of the Ferretti datasets (T. R. Ferretti, McRae, and Hatherell 2001).

Greenberg, A. B. Sayeed, and Demberg 2015 and Greenberg, Demberg, and A. Sayeed 2015 further developed the TypeDM and the role-based models, investigating the effects of verb polysemy on human thematic fit judgements and introducing a hierarchical agglomerative clustering algorithm into the prototype creation process. Their goal was to cluster together typical fillers into multiple prototypes, corresponding to different verb senses, and their results showed constant improvements of the performance of the DM-based model.

Finally, Tilk, Demberg, A. Sayeed, Klakow, and Thater 2016 presented two neural network architectures for generating probability distributions over selectional preferences for each semantic 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.

2.3. Distributional Semantics and the Problem of Compositionality

One of the most common criticisms addressed to distributional approaches is that they are often limited to the level of the content words, but they do not have any satisfactory way to represent the meaning of function words and, until recently, it did not provide a solution to the problem of the representation of more complex linguistic units, such as phrases and sentences. Research in the last decade led finally to a meeting point between DSMs and formal semantics, which has instead a strong focus on the compositional rules with which we can derive the meanings of complex expressions from the meanings of their constituents.

Some of the earliest works on compositional distributional semantics relied on sophisticated methods based either on quantum logic or on tensor products (Smolensky 1990; Clark and Pulman 2007; Widdows 2008; Rudolph and Giesbrecht 2010), but given the high complexity of the computations, they were never tested on large-scale corpus-based semantic tasks. [7]

The first research work that gained popularity to the problem of meaning composition with DSMs was by Mitchell and Lapata 2010. In their framework, these authors identified two main classes of compositional models, i.e. additive models

$$p = Au + Bv \tag{2.6}$$

where A and B are weight matrices, and multiplicative models

7. An exception is probably the early work of Van de Cruys 2010 on a non-negative matrix factorization method for the automatic induction of selectional preferences.

$$p = Cuv \tag{2.7}$$

where C is a weight tensor that projects the uv tensor product into the space of p.

From these general forms, Mitchell and Lapata derived two simplified models:

— the simplified additive model

$$p = \alpha u + \beta v \tag{2.8}$$

with both weights generally set to 1, so that the resulting vector is simply the component-wise sum of the two input vectors;

— the pointwise multiplicative model

$$p_i = u_i * v_i \tag{2.9}$$

which realizes a sort of intersection of the features of the two input vectors (all features that are not shared by both vectors are set to zero).

Despite its simplicity, this latter model turned out to be extremely effective in a wide variety of tasks and settings, including the estimation of phrasal similarity (Blacoe and Lapata 2012), the prediction of paraphrase similarity (Mitchell and Lapata 2008), paraphrase ranking (Erk and S. Padó 2008; Blacoe and Lapata 2012), and language modeling (Mitchell and Lapata 2009). [8]

8. Erk and S. Padó 2008 applied the approach of Mitchell and Lapata to the problem of computing a context-adapted meaning for verb-noun composition. Differently from the original

However, the simple models by Mitchell and Lapata have obvious drawbacks, first of all the fact of being insensitive to word order and to syntactic structure. In the attempt of dealing with such issues, some of the next works tried to generalize the simple additive model by applying structure-encoding operators to the vectors before the composition (Zanzotto, Korkontzelos, Fallucchi, and Manandhar 2010; Guevara 2010; Mitchell and Lapata 2010; Socher, E. H. Huang, Pennin, C. D. Manning, and Ng 2011).

Other approaches addressed directly the problem of compositionality in DSMs in a simil-Montagovian framework, in which meaning composition was seen essentially as a process of application of functions (Montague 1974; Baroni, Bernardi, and Zamparelli 2014).

Baroni and Zamparelli 2010 presented a study on adjectival modification in which nouns are represented as vectors, while adjectives are represented as matrices, corresponding to functions over nominal vectors induced from corpus data. The values in the adjective matrix are computed by partial least squares regression, using the co-occurrence vectors of adjectival phrases observed in corpora to train the model. The independent variables in the regression equation are the dimensions of the vectors of the component nouns, while the dimensions of the phrase vectors are the dependent variables. Coherently with the Montagovian view, which considers attributive adjectives as specific functions mapping from the meaning of a noun onto the meaning of a modified noun, each adjective corresponds to a specific matrix.

The proposal by Baroni and Zamparelli turned out to be very influential in the field, and was later extended by several researchers. E. Grefen-

model, their proposal includes also syntactic information in the form of the selectional preferences of the verbs.

stette, Dinu, Y.-Z. Zhang, Sadrzadeh, and Baroni 2013 used the regression method to represent transitive verbs by means of 3-way tensors: first, they estimate matrices of verb-object phrases from subject- and SVO-vectors; then, they estimate the verb tensors from verb-object matrices and object vectors. Bride, Van de Cruys, and Asher 2015 proposed a generalization of the model, making use of a general tensor of adjectival composition instead of multiple, single-adjective matrices, and successfully tested it on benchmarks for adjective-noun and noun-noun composition in English and French.

Other applications of this method to the cases of determiners and intransitive verbs can be found in the studies by Bernardi, Dinu, Marelli, and Baroni 2013 and Dinu, Pham, and Baroni 2013.

Still, the application of this model (known as the *lexical function model*) to real sentences is very limited, especially because of the quick growth of representation size (if noun meanings are encoded in vectors of 300 dimensions, function with two arguments like the transitive verbs will have to be represented as tensors of 300^3 dimensions). Thus, a "practical" lexical function model (henceforth PLF) has been proposed by Paperno, Pham, and Baroni 2014, who addressed the issue by representing all words with a vectors, and all functional words with one vector plus one or more matrices, one for each of their argument slots. Meanings are composed by means of two different rules:

— **function application**, applied in cases of composition of syntactic sisters with a different number of argument slots (e.g. an adjective and a noun). In such cases, the word with the higher number of slots is treated as the functor, and the other word as the argument. The composition is carried out by multiplying the argument vector by the

last matrix in the functor tuple, and then summing the result to the functor vector. Unsaturated matrices are carried up to the composed syntactic node;

— **symmetric composition** applies when two syntactic sisters have the same number of argument slots (e.g. coordination between nouns/verbs) and is carried out simply by summing the objects in the two tuples.

Thanks to this approach, Paperno and colleagues have been able to keep sentence representation to a manageable size and obtained performances competitive with the state-of-the-art on several benchmark datasets for sentence-level semantic composition. [9] A following evaluation of the PLF model was carried out by Rimell, Maillard, Polajnar, and Clark 2016, in a general comparison of compositional models on their newly-proposed benchmark, the RELPRON dataset. [10] The authors found that one version of the PLF was the best performing method, although none of the tested models was able to significantly outperform a simple vector sum.

Another related approach was the one proposed by Socher, Huval, C. D. Manning, and Ng 2012, based on rich lexical representations: each word is represented by a vector and a matrix encoding its interaction with the syntactic sisters. Compositional representations for phrases and sentences are learned by a recursive neural network model (RNN), in a supervised setting.

9. A correction of the original PLF model has been proposed more recently by Gupta, Utt, and S. Padó 2015, who noticed an inconsistency leading to an overestimation of the contribution of the predicate in the meaning of the composed phrase.

10. The dataset is composed by a set of terms and properties, and the property is expressed in the form of a relative clause, e.g. *telescope - device that astronomers use*: the goal for an efficient composition method is to generate vector representations for the relative clauses that are close in a distributional space to the term vector. The metric used for the evaluation in Rimell, Maillard, Polajnar, and Clark 2016 is Mean Average Precision, which assigns higher scores to compositional methods that are able to place a higher number of correct properties closer to each term.

However, despite successful evaluation on three different benchmarks, a problem of this method is that it assumes task-specific parameter learning, so that the results cannot be generalized across tasks.

2.3.1. Differences Between our Approach and the Lexical Function Model

Our study, as well as those presented in this subsection, is concerned on how representations for complex meanings (i.e. meanings for units beyond single words) are constructed. However, there are some important differences in our perspectives on the problem.

First, the other studies assume, in some form, the traditional principle of Fregean compositionality, as the representation of a complex units has to be derived by the representation of its immediate constituents. Although we recognize the important role played by compositionality in natural language, we do not believe it is the whole story. Rather, we think with Jackendoff that it should be conceived as an option in a wider array of alternatives (Jackendoff 1997).

Secondly, we tend to agree with the psycholinguistic theories claiming that building the semantic representations for sentences is essentially an inferential process. Sometimes, the representation of an event is already stored in the semantic memory and there is no need of *composing* it, it just needs to be *retrieved*. Sometimes, already-existing representation are "extended" by similarity with the input currently being processed.

Finally, our focus is specifically on semantic factors determining difficulty in language *comprehension*, and thus we are interested in modeling differences in processing complexity, instead of tackling more traditional

tasks such as phrase similarity and paraphrase detection.

2.4. Conclusion

In this chapter, we presented a general overview on Distributional Semantic Models, starting from the different algorithms that can be used to build the distributional representations. While illustrating the differences between structured and unstructured DSMs, we also showed the possible definitions of the context.

Then we moved to the relation between the Distributional Hypothesis and cognitive modeling, as it is possible to identify a strong version of the Distributional Hypothesis, in which word co-occurrences are directly responsible for the formation of semantic representations at the cognitive level. The efficiency of DSMs in modeling experimental results, which has been proved by several studies, is strongly supportive of the existence of such a causal relation.

Then we have summarized the literature on the DSMs for modeling the Generalized Event Knowledge, in the form of the thematic fit estimation task (McRae, M. Spivey-Knowlton, and Tanenhaus 1998; Elman 2009). Such models have been the starting point for our work on the computation of semantic complexity, as we will show in the following chapter.

Finally, we have presented the works in Distributional Semantics that have tried to deal with the compositionality issue under a formal semantics perspective, and we have discussed the fundamental differences between such an approach and our proposal.

3. Model Description and Experiments

Main contributions of the chapter:

— we introduce a Distributional Model of semantic complexity and we distinguish between a Memory and a Unification component of the model;

— the Memory component is based on the hypothesis that our semantic memory stores Generalized Event Knowledge (*GEK*);

— events in Memory are modeled by means of *syntactic joint context* directly extracted from parsed corpora;

— the Unification component consists of a function for building in the working memory a unified semantic representation. Each semantic representation is weighted in terms of both *event salience* and internal semantic coherence;

— we present two case studies for our complexity model: the composition and update of verb argument expectations on the Bicknell dataset and the modeling of logical metonymy on the datasets by McElree and Traxler.

The goal of our work is to introduce a Distributional Model for computing semantic complexity, inspired by some principles of the Memory, Unification and Control framework (Hagoort 2013; Hagoort 2016). In particular, we would like to stick with the idea that the issue about semantic compositionality is a balance between storage and computation (Baggio, Van Lambalgen, and Hagoort 2012). In comprehension, we assume the existence of a processing advantage for sentences describing typical events and situations (see the literature review in Chapter 1 for the supporting evidence).

Such a claim also requires an answer to some fundamental questions: first of all, what is the nature of the stored units, and how are they used during processing? And secondly, what kind of information could we use to model this kind of knowledge?

In brief, our model relies on the following assumptions, which will be further clarified in the next sections:

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

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

— during sentence processing, lexical items activate portions of GEK_D, which are then unified to form a coherent representation of the event expressed by the sentence. [1]

The objectives of our model are i) to build an incremental distributional representation of a sentence, and ii.) to associate a **compositional cost** to such a representation in order to account for the complexity of semantic processing.

In our view, complexity of building semantic representations depends on two main factors: a.) the availability and *salience* of "ready-to-use" event information already stored in GEK_D and b.) the cost of unifying the portions of the GEK_D activated by the context into a coherent semantic

1. In sentence comprehension there are probably many other type of cues that can activate event representations (e.g. pragmatic cues, or cues coming from the general discourse context), but for simplicity our model takes into account just the lexical items within the sentence.

representation, with the latter depending on the mutual *semantic coherence* of the event participants.

We thus predict that sentences containing highly *familiar lexical combinations* are easier to process than sentences expressing novel ones (recall the examples in Section 1.1.). Moreover, the complexity of novel combinations depends on how "compatible" they are with the event knowledge stored in the semantic memory.

After introducing our complexity model (Section 3.1), we present some experiments on compositionality-related task: the update of argument expectations in a binary classification task (Subsection 3.2.1), the modeling of processing times and the retrieval of the hidden event in logical metonymy (Subsection 3.2.2). [2]

2. Although the architecture presented here is similar to the systems described in Chersoni, Blache, and Lenci 2016 and Chersoni, Lenci, and Blache 2017, several details have been changed over the months, and thus the results are different.

3.1. A Distributional Model of Semantic Complexity

What we call Generalized Event Knowledge is knowledge about events and situations that we have experienced under different modalities, including linguistic input (i.e., the subset of GEK that we call GEK_D). GEK is assumed to be highly structured, and organized under various levels of complexity, granularity, and schematicity. It includes information about fully-specified micro-events (e.g., students read books, gardeners cut grass etc.) and about more complex scenarios (e.g. scripts and narrative schemas: see Schank and Abelson 2013).

Crucially, sentences in natural language are seen as *partial descriptions of events*. "Partial" means that many elements and details about described situations can be left unspecified, and it is up to the comprehender to infer the missing parts by retrieving the information in the GEK: for example, when we hear a sentence like *The soldier killed all the enemies*, we could infer that they used some sort of weapons (e.g. rifles, machine guns etc.) for doing that.

Each linguistic expression works as a cue for recovering portions of the GEK and, as long as comprehenders are enough to retrieve the "right" event scenario, they are also able to anticipate upcoming arguments in the sentence. 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.

Comprehension consists of recovering the most likely event expressed by a sentence (Kuperberg and Jaeger 2016), it is an incremental process

leading to the construction of a semantic representation, obtained by combining the subsets of GEK activated by the different constructions in the sentence. Analogously to Hagoort 2013 and Hagoort 2016, we distinguish between two components of our model:

— a Memory component, representing the storage of event structures in the GEK_D contained in the semantic memory;

— a Unification component, which combines the units in the GEK_D in order to obtain newer and more complex structures.

3.1.1. The Memory Component: Modeling the GEK in the 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_D, 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. Therefore, processing a linguistic expression in a given sentence will lead to the activation of a set of events in the long-term memory, each one associated to σ score.

$$\begin{bmatrix} \text{WORD} & \textit{student} \\ \text{EVENTS} & \begin{bmatrix} \text{GEK} & \langle e_1, \sigma_1 \rangle, \ldots, \langle e_n, \sigma_n \rangle \end{bmatrix} \end{bmatrix}$$

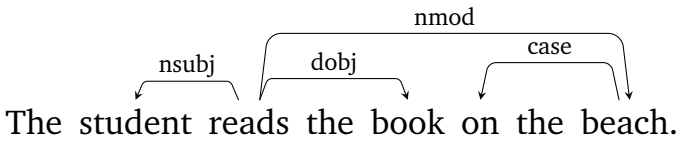
We represent **events** in GEK_D as feature structures specifying participants and roles[3], and we extract this information from parsed sentences

3. Readers that are familiar with the conventions of Sign-Based Construction Grammar (Sag 2012) will surely notice the similarities in the representation of the constructions.

in corpora: the attributes are *syntactic dependencies*[4], which we use as a surface approximation of semantic roles, and the values are *the dependent lexemes represented as distributional vectors*. The participant vectors in the event representation are meant to be "out-of-context" encoding of lexical items. Any type of distributional representation can be used to this purpose (e.g. explicit count-based vectors, low-dimensionality dense embeddings, etc.).

The type of distributional space that we use instead for representing events is a dependency-based DSM. However, as an element of novelty with respect to traditional dependency-based DSMs, we extract from corpora **syntactic joint contexts**, as defined in Chersoni, Santus, Lenci, Blache, and C.-R. Huang 2016. A syntactic joint context includes the entire set of dependencies of a given lexical head (ignoring determiners and modifiers), and we assume it as surface representations of events.[5]

For example, from the sentence *The student reads a book* we extract the following event representation:



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

5. Of course, using dependencies instead of semantic roles is a rough approximation, especially for those roles (e.g. instruments, locations) that are typically expressed in syntax by prepositional complements. However, recent evaluations carried out on dependency- and semantic role-based spaces on thematic fit modeling tasks suggest that in practice it works reasonably well (A. Sayeed, Demberg, and Shkadzko 2015; Santus, Chersoni, Lenci, and Blache 2017).

$$\begin{bmatrix} \textbf{EVENT} & \\ \text{NSUBJ} & \overrightarrow{student} \\ \text{HEAD} & \overrightarrow{read} \\ \text{DOBJ} & \overrightarrow{book} \\ \text{NMOD-ON} & \overrightarrow{beach} \end{bmatrix}$$

Events in GEK_D can be cued by several lexical items, with a strength depending on the salience of the event given the item. In the following example, the same event is cued by *student*, *read* and *book* (we represent the cue-event activation relationship by means of re-entrance).

(1) $$\begin{bmatrix} \textit{student}\ \boxed{1} & \\ \begin{bmatrix} \textbf{EVENT} & \\ \text{NSUBJ} & \boxed{1} \\ \text{HEAD} & \overrightarrow{read} \\ \text{DOBJ} & \overrightarrow{book} \\ \text{NMOD-IN} & \overrightarrow{library} \end{bmatrix} \end{bmatrix}$$

(2) $$\begin{bmatrix} \textit{read}\ \boxed{1} & \\ \begin{bmatrix} \textbf{EVENT} & \\ \text{NSUBJ} & \overrightarrow{student} \\ \text{HEAD} & \boxed{1} \\ \text{DOBJ} & \overrightarrow{book} \\ \text{NMOD-IN} & \overrightarrow{library} \end{bmatrix} \end{bmatrix}$$

(3)
$$\begin{bmatrix} \textit{book}\ \boxed{1} \\ \begin{bmatrix} \textbf{EVENT} & \\ \text{NSUBJ} & \overrightarrow{student} \\ \text{HEAD} & \overrightarrow{read} \\ \text{DOBJ} & \boxed{1} \\ \text{NMOD-IN} & \overrightarrow{library} \end{bmatrix} \end{bmatrix}$$

Besides complete events, we assume GEK_D to contain schematic (i.e., underspecified) events too. For instance, from the sentence *The student reads a book* we also generate **schematic events**.

$$\begin{bmatrix} \textbf{EVENT} & \\ \text{NSUBJ} & \overrightarrow{student} \\ \text{DOBJ} & \overrightarrow{book} \\ & \end{bmatrix}$$

Schematic events can be 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_D is not a flat list of events, but a structured repository of prototypical knowledge about event contingencies.

$$
\begin{array}{ccc}
 & \begin{bmatrix} \textbf{EVENT} & \\ \text{NSUBJ} & \overrightarrow{student} \\ \text{DOBJ} & \overrightarrow{book} \end{bmatrix} & \\
\begin{bmatrix} \textbf{EVENT} & \\ \text{NSUBJ} & \overrightarrow{student} \\ \text{HEAD} & \overrightarrow{read} \\ \text{DOBJ} & \overrightarrow{book} \end{bmatrix} & \begin{bmatrix} \textbf{EVENT} & \\ \text{NSUBJ} & \overrightarrow{student} \\ \text{HEAD} & \overrightarrow{buy} \\ \text{DOBJ} & \overrightarrow{book} \end{bmatrix} & \begin{bmatrix} \textbf{EVENT} & \\ \text{NSUBJ} & \overrightarrow{student} \\ \text{HEAD} & \overrightarrow{write} \\ \text{DOBJ} & \overrightarrow{book} \end{bmatrix}
\end{array}
\tag{4}
$$

It is worth remarking that the events in GEK_D are complex symbolic structures including distributional representations of the event head and its participants. Events in GEK_D are therefore modeled as a sort of **semantic frames** whose elements are distributional vectors. However, there is an important difference: unlike traditional semantic frames, our events are saturated structures, with already-specified participants for each role.

3.1.2. 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; Zwaan and Madden 2005). Comprehension always occurs within the context of

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

an existing SM: during online sentence processing, lexical items cue portions of GEK_D and the SM is dynamically updated by unifying its 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 (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_D associated with w_{i+1} in the lexicon, $GEK_D[w_{i+1}]$, is activated;
2. $GEK_D[w_{i+1}]$ 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_D cued by the linguistic elements that are composed:

$$F(SM_i, GEK_D[w_{i+1}]) = SM_{i+1} \tag{3.1}$$

The composition function is responsible for two distinct processes:

— F **unifies** two event feature structures into a new event, provided that

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).

the attribute-value features of the input events are compatible. Here is an example of unification:

$$\begin{bmatrix} \textbf{EVENT} & \\ \text{NSUBJ} & \overrightarrow{student} \\ \text{DOBJ} & \overrightarrow{book} \end{bmatrix} \bigsqcup \begin{bmatrix} \textbf{EVENT} & \\ \text{NSUBJ} & \overrightarrow{student} \\ \text{HEAD} & \overrightarrow{read} \end{bmatrix} = \begin{bmatrix} \textbf{EVENT} & \\ \text{NSUBJ} & \overrightarrow{student} \\ \text{HEAD} & \overrightarrow{read} \\ \text{DOBJ} & \overrightarrow{book} \end{bmatrix} \tag{3.2}$$

This is instead an example of failed unification:

$$\begin{bmatrix} \textbf{EVENT} & \\ \text{NSUBJ} & \overrightarrow{student} \\ \text{DOBJ} & \overrightarrow{beer} \end{bmatrix} \bigsqcup \begin{bmatrix} \textbf{EVENT} & \\ \text{NSUBJ} & \overrightarrow{student} \\ \text{HEAD} & \overrightarrow{read} \\ \text{DOBJ} & \overrightarrow{book} \end{bmatrix} = FAIL \tag{3.3}$$

In the first case, the event of a *student* performing an action on an *book* and the event of a *student reading* something are unified into a new event of a *student reading a book*; in the second case, the two event structures cannot be unified, because of the clash between the values of the DOBJ attribute. The updates of the SM, as new words are processed, will progressively rule out the activated events that fail to satisfy the compatibility constraints;

— 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 *global semantic coherence* and its activation

by lexical items or *salience*.

The two above-mentioned factors, which will be introduced in the following subsections, are essential in our account for the complexity of semantic representations.

3.1.3. The Cost of Unification: Semantic Coherence

We use a θ_{e_k} score 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:

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

The event represented in (5-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 (5-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.

Verb-argument typicality is measured in the computational and psycholinguistic literature with *thematic fit* values (McRae, M. Spivey-Knowlton, and Tanenhaus 1998; see Section 1.2). In the present proposal, the notion of thematic fit is extended in order to account for the degree of coherence

of events described by natural language sentences in their entirety.

In computational approaches (Erk, S. Padó, and U. Padó 2010,Baroni and Lenci 2010), thematic fit is assessed by means of vector cosine in the following way:

> $\theta(\overrightarrow{a}|s_i, \overrightarrow{b})$ (the thematic fit of the lexical item $\overrightarrow{a}$ given the lexical item $\overrightarrow{b}$ 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

For instance, the thematic fit of *student* as an agent in *writing*-events is given by the cosine between the vector 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 vector 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 vector representation and the ideal filler in a DSM.

Although we adopt the same approach for measuring the typicality of the participants, an important problem for us is, *how the partial scores of single event-participant combinations have to be combined in a global semantic coherence score?*

In our work, we experimented with two different solutions:

— as in Chersoni, Blache, and Lenci 2016 and Chersoni, Lenci, and Blache 2017, 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*[8];

— similarly to Lenci 2011 and to Chersoni, Santus, Blache, and Lenci 2017, 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 (described by the sentence se_k) is defined as:

$$\theta_{e_k} = \prod_{a,b,s_i \in e} \theta(\overrightarrow{a}|s_i, \overrightarrow{b}) \tag{3.4}$$

For example, given a sentence like *The student drinks wine*, the score θ_{e_k} would be the product of three factors:

— the thematic fit of *student* as an *agent* of *drink*;

— the thematic fit of *beer* as a *co-participant* of *student*;

— the thematic fit of *beer* as a *patient* of *drink*.

Thus, θ_{e_k} would be computed as:

8. 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 also due to inter-arguments typicality (e.g. the facilitation for sentences with typical agent-patient combinations in Bicknell, Elman, Hare, McRae, and Kutas 2010).

$$\theta_{e_k} = \theta(\overrightarrow{student}|agent, \overrightarrow{drink}) * \theta(\overrightarrow{beer}|coParticipant, \overrightarrow{student}) * \theta(\overrightarrow{beer}|patient, \overrightarrow{drink}) \quad (3.5)$$

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. While presenting the results, we will refer to the models holding this interpretation of the semantic coherence score as $ThetaProd$.

The alternative approach consists of 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 the verb given the previously-realized arguments.
In this model, the update on the expectation EX for a given filler caused by new input (e.g., a verb combining with an agent) is modeled by means of a function $f(x)$ that combines the prototypes built out of the typical fillers for every $input$ word.

$$EX_{role}(\langle input_1, input_2 \rangle) = f(EX_{role_1}(input_1), EX_{role_2}(input_2)) \quad (3.6)$$

Once the expectation vector has been calculated, the $filler$ fit for a $role$, given $\langle input1, input2 \rangle$, can be computed by measuring the cosine similarity between the $filler$ and the expectations vector.

As an example, if we wanted to estimate how likely is *burglar* as a patient of *the policeman arrested the...*, the procedure would be the following:

1. we first build a prototype out of the vectors of nouns typically co-occurring with the agent *policeman-n*;
2. then we build another prototype for the vectors of typical patients of the verb *arrest-v*;
3. we combine the prototype vectors through $f(x)$;
4. at this point, we can estimate the $filler$ fit by calculating its cosine similarity to the resulting prototype.

$$EX_{patient}(burglar|\langle police, arrest\rangle) = \\ cosSim(burglar, f(EX_{coParticipant}(policeman), EX_{patient}(arrest))) \quad (3.7)$$

In Chersoni, Santus, Blache, and Lenci 2017, the best performing function turned out to be the simple vector sum between prototype vectors, and thus we used vector sum for the experiments presented in the following chapter.

$$EX_{patient}(patient|\langle agent, verb\rangle) = \\ cosSim(patient, sum(EX_{cooc\_patient}(agent), EX_{patient}(agent))) \quad (3.8)$$

This model captures a different kind of intuition, in the sense that semantic coherence is conceived as the coherence between the dynamically-updated expectations for the participants of an event described by a sen-

tence, on the one hand, and the fillers saturating the participant roles on the other hand. The global semantic coherence corresponds to the typicality of the last argument to be predicted in the sentence.

$$\theta_{e_k} = EX_{lastRole} \tag{3.9}$$

In our experiments, we will refer to this model as $ThetaProtoSum$.
It should be noticed that both models compute thematic fit scores as cosine similarity scores between fillers and prototypes. Although this is the most common approach to the thematic fit task, it is not the only one: for example, Santus, Chersoni, Lenci, and Blache 2017 recently showed that a metric based on weighted feature overlap outperforms vector cosine in many settings. However, this approach requires the estimation of an additional parameter (i.e., the extent of the overlap to be taken into account), so we have decided to adopt the standard cosine metric.

3.1.4. 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 this process, 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 the GEK is computed by summing the activation scores of the single lexical items

cuing it (i.e., the corpus-estimated conditional probabilities of the event given each lexical item in the input sentence):

$$\sigma_i = P(e|i) = \frac{P(e,i)}{P(i)} \tag{3.10}$$

$$F(\sigma_i, \sigma_j) = \sigma_{e_k} = \sigma_i + \sigma_j \tag{3.11}$$

Thus, 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 should incrementally increase their salience.

It should be pointed out that the activation mechanism does not work only for fully-saturated 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 the GEK (otherwise the contribution of this component will be 0); 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:

$$\sigma_e = \sum_{e_i \in E} \sigma_{e_i} \tag{3.12}$$

where the set of events E includes both the full event e_k and all the sub-events activated by the lexical items in s_{e_k}.

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_D and retrieved during sentence processing.
Salience scores can also be used to identify missing pieces of information, such as implicit arguments. For instance, suppose 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* becomes a highly salient (i.e., highly probable) location for the event described in the sentence. This is a piece of unexpressed information, which is recovered from the SM built during sentence comprehension.

The θ score, on the other hand, allows us to weight 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_{e_k}, its interpretation INT(s) is the event e_k with the highest **semantic composition weight (SCW)**, defined as follows:

$$\text{INT}(s) = \underset{e}{\text{argmax}}(\text{SCW}(e)) \tag{3.13}$$

$$\text{SCW}(e) = \theta_e + \sigma_e \tag{3.14}$$

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

$$\text{SemComp}(s) = \frac{1}{\text{SCW}(\text{INT}(s))} \tag{3.15}$$

In other words, 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. 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 events described in natural language. As shown in the next section, we tested our framework on some psycholinguistics datasets, in the attempt of modeling the complexity-related effects of the studies of reference.

3.2. Experimental Settings

In order to carry out our experiments, the first step was to populate our GEK_D with events extracted from parsed corpora. We followed the procedure proposed by Chersoni, Santus, Lenci, Blache, and C.-R. Huang 2016 to extract syntactic joint contexts from a concatenation of four different corpora: the Reuters Corpus Vol.1 (D. D. Lewis, Yang, Rose, and Li 2004); the Ukwac and the Wackypedia Corpus (Baroni, Bernardini, Ferraresi, and Zanchetta 2009) and the British National Corpus (Leech 1992). 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 (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. As in Chersoni, Santus, Lenci, Blache, and C.-R. Huang 2016, we discarded all the adjec-

tival/adverbial modifiers and just kept 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 GEK_D. 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 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 have been weighted by means of Local Mutual Information (Evert 2004):

$$LMI(t, r, f) = log\left(\frac{O_{trf}}{E_{trf}}\right) * O_{trf} \tag{3.16}$$

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. 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):

— the SUBJ relation for the agent role;

— the DOBJ relation for the patient role;

— a generic VERB relation for co-participants. Concretely, this relation links noun pairs that appear as subject and direct object of the same verb. [9]

3.2.1. Composing and Updating Verb Argument Expectations

As a first test for our framework, we measure the semantic complexity of the sentences in the Bicknell dataset (Bicknell, Elman, Hare, McRae, and Kutas 2010). The Bicknell dataset was prepared to verify the hypothesis that the typicality of a verb direct object depends on the subject argument. For this purpose, the authors selected 50 verbs, each paired with two agent nouns that altered the scenario evoked by the agent-verb combination.

Plausible patients for each agent-verb pair were obtained by means of production norms, in order to generate triplets where the patient was congruent with the agent and with the verb. For each congruent triple, they also generated an incongruent triple, by combining each verb-congruent patient pair with the other agent noun, in order to have items describing atypical situations.

The final dataset included 100 pairs agent-verb-patient triplets, that were used to build the sentences for a self-paced reading and for an ERP experiment. [10]

To give an example, experimental subjects were presented with sentence

9. An analogous relation had been used by Lenci 2011 in his study on the composition of argument expectations.

10. Actually, Bicknell and colleagues used only a subset of 64 pairs, after removing the items that were potentially problematic for their experiments (i.e. cases in which the choice of the agent argument does not lead to noticeable changes in the expectations for the patient). In our experiments, we used instead the original dataset.

pairs such as:

— The **journalist checked** the **spelling** of his latest report. (*congruent condition*)

— 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, whereas it is an implausible filler in the incongruent condition. Bicknell, Elman, Hare, McRae, and Kutas 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.

Lenci 2011 evaluated a model for composing and updating argument expectations on the ability to assign a higher thematic fit score to the congruent combinations than to the incongruent ones. For the purpose of distributional modeling, the input of Lenci's DSM were the agent-verb-patient triplets, that had been used to build the sentences of the original experiment, e.g. the input triple for the sentence *The journalist checked the spelling of his latest report* is *journalist check spelling* (labeled with syntactic relations used as proxy: other examples are shown in Table 3.1).

We interpret Bicknell's experimental data as suggesting that congruent sentences are less semantically complex than incongruent sentences. Consistently, we predict that our models will assign a higher semantic complexity score to incongruent triplets than to congruent ones. For our evaluation,

Triple	Condition
SUBJ:general HEAD:assemble OBJ:troop	congruent
SUBJ:craftsman HEAD:assemble OBJ:troop	incongruent
SUBJ:composer HEAD:arrange OBJ:music	congruent
SUBJ:librarian HEAD:arrange OBJ:music	incongruent

Table 3.1. – Examples of the triplets of the Bicknell dataset in the congruent and the incongruent condition.

we score a hit for each time that, given a congruent-incongruent triple pair, a model assigns a higher $SemComp$ score to the incongruent one. Models are primarily evaluated in terms of their accuracy in this binary classification task.

3.2.1.1. Complexity Models

For each test triple, we had to compute a σ and a θ score:

— θ represents the semantic coherence of the event represented by the sentence, and it is obtained by measuring the mutual typicality of its components. As we anticipated in Subsection 3.1.3, we built two models that differ for the way of assessing semantic coherence:

1. in the first model, $ThetaProd$, we computed the θ values as the product of partial thematic fit scores. Following equation 3.4, we compute θ_e for each triple as the product of i) the thematic fit of SUBJ 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 SUBJ, $\theta_{S,O}$.

   In particular, $\theta_{S,V}$ is the cosine between the vector of SUBJ and the centroid vector built out of the k most salient subjects of the verb HEAD (e.g., the cosine between the vector of *general* and

the centroid vector of the most salient subjects of *assemble*); $\theta_{O,V}$ 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 *troop* and the centroid vector of the most salient objects of *assemble*); and $\theta_{O,V}$ is the cosine between the vector of DOBJ and the centroid vector built out of the k most salient direct objects occurring in events whose subject is SUBJ (e.g., the cosine between the vector of *troop* and the prototype vector of the most salient objects of events whose subject is *general*);

2. in the second model, $ThetaProtoSum$, the θ_e of each triple was computed as the similarity score between the vector of the patient and the vector of the expectations for the patient given the agent and the verb, as in Equation 3.8. (since we are using syntactic approximations of roles, $agent = SUBJ, verb = HEAD, patient = DOBJ$). Simple vector sum was the function that we used to combine partial prototypes in the global expectation vector for the patient (Chersoni, Santus, Blache, and Lenci 2017).

For both our models, after using Local Mutual Information to weight the dependency-based space, we extracted the typical fillers for each role as the nouns f with the strongest LMI association with the target word t and the relation r (the syntactic proxy of the target role). Following Baroni and Lenci 2010, we set the parameter k (i.e. the number of typical fillers used to build the prototypes) to 20.

— as for the σ score in all models, given an event e_k, we look for a matching syntactic joint context in our GEK_D repository and for schematic

events matching the sub-chunks of e_k (some examples are shown in Table 3.2.). For each of these events $e_i \in E$, if they are present in GEK_D, we compute an activation score by using Equations 3.10 and 3.11. Partial scores are summed with Equation 3.12 to obtain the global σ_e.

Syntactic Joint Context	Schematic Events
SUBJ:general HEAD:assemble OBJ:troop	(SUBJ:general HEAD:assemble) (HEAD:assemble OBJ:troop) (SUBJ:general OBJ:troop)
SUBJ:journalist HEAD:write OBJ:article	(SUBJ:journalist HEAD:write) (HEAD:write OBJ:article) (SUBJ:journalist OBJ:article)

Table 3.2. – Examples of the the schematic events to be retrieved in the GEK for computing the σ of a given joint context.

Finally, once we have computed θ_e and σ_e for each of our test triplets, we used Equation 3.14 and 3.15 to derive the final $SemComp$ scores.

3.2.1.2. Baseline Models

Apart from our models of semantic complexity, we prepared two baselines inspired by the early models of distributional compositional semantics by Mitchell and Lapata 2010.

As described in chapter 2.3, Mitchell and Lapata proposed two simple models for vector composition:

— the simplified additive model (SUM)

$$p = \alpha u + \beta v \tag{3.17}$$

where both weights are set to 1 (the output vector is the component-

wise sum of the input ones);

— the pointwise multiplicative model ($PRODUCT$)

$$p_i = u_i * v_i \tag{3.18}$$

Despite their simplicity, such models turned out to be extremely efficient and competitive in a wide variety of compositionality-related tasks (Mitchell and Lapata 2010; Rimell, Maillard, Polajnar, and Clark 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, either by summing or by multiplying the respective vectors. 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 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 the other argument.

Another baseline model is based on the notion of *Surprisal*. After extracting all the subject-verb-object triples from our training corpora, 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:

$$Surprisal(t) = -\log_2 P(subject(t), verb(t), object(t)|subject(t), verb(t)) \tag{3.19}$$

where $subject(t)$, $verb(t)$ and $object(t)$ are, respectively, the agent, the verb and the patient of t. For this model, the accuracy is computed as the percentage of atypical triples to which it assigns a higher surprisal score.

For each dataset, we will also present a comparison with the models introduced in the previous literature. In the case of the Bicknell dataset, we compare the accuracy scores to those obtained by the best system by Lenci 2011, i.e. 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 most typical fillers in the updated ranking. In this way, arguments that are not compatible with *both* the verb and the agent are filtered out.

3.2.1.3. Results on the Bicknell dataset

First of all, all the models except for the Sum baseline are able to differentiate between the two conditions. We run the Wilcoxon rank sum test on the output scores of the different models and we found that:

— the $SemComp$ scores assigned by $ThetaProtoSum$ to the incongruent condition are significantly higher ($p < 0.05$);

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

Table 3.3. – Accuracy scores plus coverage for each model on the classification task on the Bicknell dataset.

— the $SemComp$ scores assigned by $ThetaProd$ to the incongruent condition are significantly higher ($p < 0.01$);

— the thematic fit scores assigned by the baseline $Prod$ to the incongruent condition are significantly lower ($p < 0.01$). [11]

Perhaps surprisingly, the simple $Product$ baseline manages to obtain the best accuracy in the binary classification task and to discriminate between the two experimental conditions. This confirms that it is difficult to beat baselines based on simple vector operations in many compositionality-related tasks, as shown by the results by Mitchell and Lapata 2010 and Rimell, Maillard, Polajnar, and Clark 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 perform at high levels of accuracy and assign significantly different scores to the two conditions.

We consider the performance of $ThetaProd$ to be particularly satisfac-

11. The scores of the baselines are not reversed as the $SemComp$ 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.

tory, as it manages to outperform the original model of expectations update by Lenci 2011, when tested on the covered triplets (73.8%) [12]. 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, $ThetaProtoSum$ ($p < 0.1$). [13] Compared to the other baselines, its advantage over Sum is significant at $p < 0.05$, while the one over the $Surprisal$ baseline is only marginally significant ($p < 0.1$).

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 0, as no matching joint context was retrieved from the GEK. 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, Santus, Blache, and Lenci 2017) corresponding to our θ component. Thus, one might wonder if it is actually worth making the model more complex by introducing, with σ, an extra parameter.

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

Table 3.4. – Accuracy scores for the two complexity models without the σ component and the accuracy loss with respect to the full model.

Table 3.4 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

12. The accuracy score has been provided by the author himself.
13. p-values have been computed with the $\tilde{\chi}^2$ test.

GEK 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) were too small to reach significance, the contribution of the two components seems to be more balanced in $ThetaProd$.

From these data, it seems anyway 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, independently from the smoothing), and the future developments of complexity model will have to take this factor into account. Such a problem could me mitigated by introducing a robust generalization component, which could generating new joint contexts by making inferences on new potential event participants, or by integrating the model with a system that is able to generate event representations from other types of data (e.g. images).

3.2.2. Distributional Modeling of Logical Metonymies

The interpretation of the so-called *logical metonymies* (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*).[14] Logical metonymy is a clear case of what Jackendoff 1997 calls "enriched composition", as opposed to Fregean compositionality. In Fregean compositionality, the elements of content in the meaning of a sentence are found in the lexical conceptual structure of the lexical items composing it, and the ways in which lexical meanings are combined is entirely guided by the syntactic structure of the sentence. In enriched composition instead, the conceptual structure of a sentence may contain material that is not overtly expressed, but that must be present in the conceptual structure in order to satisfy a well-formedness requirement (as in the case of logical metonymies, in which the type clash needs to be "repaired") or other constraints coming from the pragmatics of the discourse or from the extralinguistic context.

Past research work brought extensive evidence that such metonymic constructions also determine extra processing costs during online sentence comprehension (McElree, Traxler, Pickering, Seely, and Jackendoff 2001; Traxler, Pickering, and McElree 2002), although such evidence is not un-

14. It is important to point out that the present part of the chapter is not concerned with metonymy as a general linguistic operation (e.g. we do not consider cases such as *wanting a beer*), but only with the more circumscribed phenomenon of logical metonymy, as defined in the theoretical and in the experimental literature on sentence processing.

controversial (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, Pickering, and McElree 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, S. Padó, and Lenci 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. [15]

15. Consider the classical example from Lascarides and Copestake 1998: *My goat eats anything. He really enjoys 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.

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 Almeida 2004, 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 S. 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 S. Padó 2011 and Zarcone, S. Padó, and Lenci 2014 brought experimental evidence for the role of Generalized Event Knowledge (GEK) in the interpretation of logical metonymies.[16] The experiments on German by Zarcone, S. Padó, and Lenci 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 Elman's words-as-cues paradigm (Elman 2009; Elman 2014): since the words in the mental lexicon are cues to event knowledge, the type clash between verb and direct

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

object is solved by retrieving the hidden activity that is more typical, given the context.

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, Keller, and Scheepers 2003; Shutova 2009) or of distributional similarity-based thematic fit estimations (Zarcone, Utt, and S. Padó 2012), whereas the second aspect concerns modeling the experimental data about processing costs. Zarcone, Lenci, S. Padó, and Utt 2013 showed that a distributional model of verb-object thematic fit can reproduce the reading times differences in the experimental conditions found by McElree, Traxler, Pickering, Seely, and Jackendoff 2001 and Traxler, Pickering, and McElree 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. Therefore, our *SemComp* weights will be used to carry out two different tasks:

— modeling the reading times of metonymic sentences, as reported in previous experimental studies;

— the prediction of the covert event in a binary classification task.

3.2.2.1. Logical Metonymy: Experiments and Datasets

We applied our models of semantic complexity to account for psycholinguistic data about metonymic sentences. In particular, we predicted that *metonymic sentences 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, Lenci, S. Padó, and Utt 2013,

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.

Additionally, we predicted 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 are overtly realized. The idea, in other words, is that the inferred covert event in logical metonymy is *the one maximizing the Semantic Composition Weight (SCW) of the global event structure* that represents the interpretation of the sentence.

For our experiments, we used two datasets created for previous psycholinguistic studies: the McElree dataset (McElree, Traxler, Pickering, Seely, and Jackendoff 2001) and the Traxler dataset (Traxler, Pickering, and McElree 2002).

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

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

Sentence (6-a) corresponds to the metonymic condition (MET), while sentences (6-b) and (6-c) correspond to non-metonymic constructions, with the difference that (6-b) represents a typical event given the subject and the object (HIGH_TYP), whereas (6-c) expresses a plausible but less typical event (LOW_TYP).

The McElree dataset was created for the self-paced reading study by McElree, Traxler, Pickering, Seely, and Jackendoff 2001, and includes 99 sentences (33 triplets), while the Traxler dataset was used in the eye-tracking experiment by Traxler, Pickering, and McElree 2002 and contains 108 sentences (36 triplets).

3.2.2.2. Logical Metonymy: Test Settings

As for the GEK_D repository and the dependency-based DSM in which the single dataset words are represented, we used the same of the previous experiment. Three triplets of the McElree datasets were discarded, since we had no coverage for them (i.e. some of the words in the triplets had very low frequency in the training corpora).

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

Table 3.5. – Coverage for the evaluation triplets

Again, the models have been tested on the triplets corresponding to the agent-verb-patient combination of the original datasets and, as for the task of modeling processing times, σ and θ scores have been computed exactly in the same way. The evaluation, in this first case, consists simply in the comparison between the scores obtained by our models ($ThetaProd$ and $ThetaProtoSum$, plus the baselines Sum and $Product$) and the outcome of the experimental studies by McElree, Traxler, Pickering, Seely, and Jackendoff 2001 and Traxler, Pickering, and McElree 2002.

Concerning the identification of the covert event, things are slightly different. Our assumption was that the interpretation of a metonymic sen-

tence like *The author starts the book* is the following conjunction of events:

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

$[_{EVENT}$ NSUBJ:$\overrightarrow{author}$ HEAD:$\overrightarrow{e}$ DOBJ:$\overrightarrow{book}]$

where e is the covert event to be recovered (e.g., *writing*).

$$\begin{bmatrix} \textbf{EVENT} & \\ \text{NSUBJ} & \overrightarrow{author} \\ \text{HEAD} & \overrightarrow{begin} \\ \text{DOBJ} & \begin{bmatrix} \textbf{EVENT} & \\ \text{HEAD} & \overrightarrow{write} \\ \text{DOBJ} & \overrightarrow{book} \end{bmatrix} \end{bmatrix}$$

We modeled covert event retrieval as a binary classification task: remember that, given a set of candidate hidden events, we argued that the selected interpretation is the one that minimizes complexity.

To test our claim, we used the following procedure inherited from Zarcone, Utt, and S. Padó 2012:

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:

   — author start write book (HIGH_TYP_EVENT)

   — author start read book (LOW_TYP_EVENT)

2. for each sentence $SV_{met}O$, we computed SCW(e) (cf. equation 3.14) of the events composing its interpretation, that is $[_{EVENT}$ S V_{met} $E_{cov}]$ and $[_{EVENT}$ S E_{cov} O] (i.e. given a quadruple pair, we computed it for both the HIGH_TYP and the LOW_TYP quadruple); [17]

3. 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).

Concerning the baselines, apart from those that were previously introduced, we also compare our results with the following systems:

1. for the task of modeling processing times

   — Zarcone, Lenci, S. Padó, and Utt 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

   $$1 - thematicFit(patient) \tag{3.20}$$

   . The model, which will be referred to as ZETAL13, does not take into account the agent filler.

   — a second surprisal model, similar to the one described in the recent study by Delogu, Crocker, and Drenhaus 2017 on logical metonymy (SURPRISALD17). In this model, we compute the probabilities of the corpora-extracted trigrams composed by the verb, a determiner and the object noun (for simplicity, we abstract away from the determiner). Given a trigram t, its surprisal score is computed as follows:

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

$$SurprisalD17(t) = -\log_2 P(verb(t), DET, object(t)|verb(t), DET, object(t)) \tag{3.21}$$

where $verb(t)$ and $object(t)$ are, respectively, the verb and the patient of the triple t, and DET is a generic determiner, whose head is the object. In their eye-tracking and ERP experiments, Delogu and colleagues 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;

2. for the event retrieval task, we used as comparison the best probabilistic model by Zarcone, Utt, and S. Padó 2012 and we compute the probability P_e of a candidate verb as the hidden event e as:

$$P(e) = P(verb) * P(subject|verb) * P(object|verb) \tag{3.22}$$

The model by Zarcone and colleagues (ZETAL12) is a generative one, in the sense that it first assumes a hidden event e and then generates the arguments on the basis of the choice of e. In a comparison with distributional models of logical metoymy, ZETAL12 obtained the highest accuracy results, to the cost of lower coverage (due to the zero-counts of many of the co-occurrences needed for probability estimation).

3.2.2.3. Modeling the Processing Times of Logical Metonymies: Results

In the task of modeling the processing times, the $ThetaProd$ model turns out to be most faithful to the results of the studies of reference.

On the McElree dataset, the Kruskal-Wallis rank sum test revealed a main effect of the sentence types on the $SemComp$ scores assigned by $ThetaProd$ ($\chi^2 = 17.18$, $p < 0.001$). 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, Traxler, Pickering, Seely, and Jackendoff 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 3.6. – 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 $ThetaProd$.

A main effect of sentence types on the $SemComp$ scores also was found for the Traxler dataset ($\chi^2 = 15.39$, $p < 0.001$). In their eye-tracking experiment (Experiment 1), Traxler, Pickering, and McElree 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: 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 ($p < 0.05$).

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

Table 3.7. – 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 $ThetaProd$.

The differences between the scores found by $ThetaProd$ are shown in Tables 3.5 and 3.5, and the comparison between conditions is made visible in Figure 3.1.

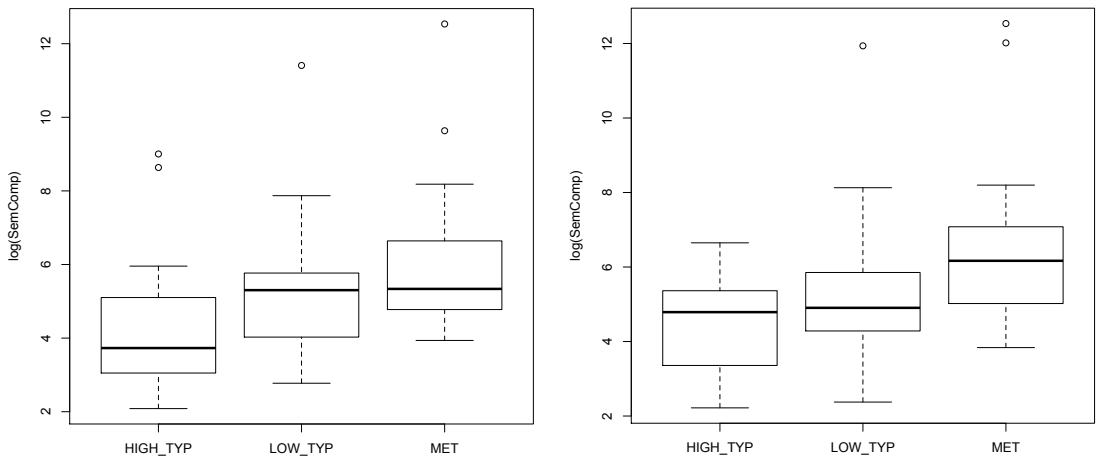


Figure 3.1. – $SemComp$ scores for McElree (left) and Traxler (right), computed with the $ThetaProd$ model.

Also the $ThetaProtoSum$ model 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 3.7 and 3.8). This pattern is very close to the one found by ZETAL13, which discrimi-

nates 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.

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

Table 3.8. – 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	0.68	-
MET	2.4e-07*	0.00084*

Table 3.9. – 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 $ThetaProtoSum$.

Concerning the surprisal 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 DSM-based 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 significance pattern is different than the one reported by McElree and colleagues, since no significant difference

between HIGH_TYP and LOW_TYP has been found ($p = 0.9$).

3.2.2.4. Identifying the Covert Event: Results

The results for the covert event identification are shown in Table 3.10. Overall, it can be observed 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 (p-values have been computed with the χ^2 test; but also due to the small size of the datasets, the differences between the models are rarely significant). 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.

Model	McElree	Traxler
Random	50% (30)	50% (36)
Sum	40% (30)	50% (36)
$Prod$	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 3.10. – Accuracy (and coverage) of the models and of the baselines on the binary classification task for covert event retrieval.

The model that goes closer to $ThetaProd$ in terms of accuracy is the reimplementation of ZETAL12. As it was found 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 also tested the $ThetaProd$ model by removing the σ component, in order to assess its contribution to the classification task. The results are shown in Table 3.11.

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

Table 3.11. – Accuracy of $ThetaProd$ after the removal of σ and performance drop on the McElree and the Traxler datasets.

Once again, the contribution of the σ component is limited to few triplets, especially on the Traxler dataset that includes several rare words. 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 GEK_D.

As a final experiment, we wanted to test the claim by Zarcone, Lenci, S. Padó, and Utt 2013, 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 could be a consequence of the dispreference of the verb for the entity-denoting argument. This equals to say, 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. [18]

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.

18. 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 $1/n$.

3.3. Conclusions

In this chapter, we have introduced our framework for the computation of semantic complexity, relying on the two components of Memory and Unification. The first refers to the storage of GEK events (that we represent by means of syntactic joint contexts), whereas the second concerns the constraint-driven combination of the units stored in the GEK into more complex structures.

The general idea at the core of our DSM is that words work as cues to the GEK, and that the recovered fragments are dynamically unified to build a representation of the events that natural language sentences are likely to communicate. Event representations are weighted along two dimensions:

— the **semantic coherence** θ of the unified event, which depends on the mutual typicality between the participants and is assessed by means of a distributional model of thematic fit;

— the **activation by lexical items or salience** σ, i.e. the strength of activation of the GEK events that are cued by the words in the sentence. 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.

An important assumption is that the complexity of building semantic representations is inversely-related to these factors: a sentence is easier to process i) when it strongly activates a corresponding event in the GEK; ii) when its participants are mutually typical, thus describing a more predictable situation.

We also presented the results of our experiments, in which we compared the predictions of our model with the findings of some psycholinguistic

studies. A model version that turned out to be particularly successful was the so-called $ThetaProd$, which computed the θ as the product of the single event-participant thematic fit scores.

Among the positives of this model:

— it was able to achieve a competitive performance on the binary classification task for the update of context-sensitive argument typicality, with the second-best scores after a very strong $Product$ baseline (which is, on the other hand, outperformed in the other tasks);

— on the logical metonymy datasets from the sentence processing studies by McElree, Traxler, Pickering, Seely, and Jackendoff 2001 and Traxler, Pickering, and McElree 2002, it was the only model to reproduce exactly the same significance patterns reported for the three experimental conditions (typical, non-typical and metonymic event);

— it achieves a remarkable accuracy on the event retrieval task, which has also been framed as a binary classification task; noticeably, in a task that turned out to be extremely difficult for all our models, it was the only one to beat significantly a random baseline. Compared to its closest competitor, a probabilistic model by Zarcone, Utt, and S. Padó 2012, it is more robust to data sparsity;

— comparing metonymic events and the corresponding structure after the integration of the covert event, the θ component assigns significantly different scores. This is coherent with the hypothesis of Zarcone, Lenci, S. Padó, and Utt 2013, according to which the event retrieval process 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;

— 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 is somehow complementary.

In general, the coverage of the σ component was found to be low on all datasets. On the one hand, this makes sense to us, since it is difficult to think that a semantic memory component could store all possible events: in most cases, it is likely that the semantic representation of the situation being communicated has to be built from scratch. On the other hand, it is desirable that future extensions of such a model implement some sort of similarity-based generalization, in order to increase the contribution of σ.

4. Conclusion

In the previous chapter, we have introduced a theoretical and computational framework for computing the semantic complexity of simple sentences, conceived as related to extra processing difficulty, and we tested it on some datasets derived from the psycholinguistic literature in order to compare the findings.

Our framework, based on the idea of a division of labor between a Memory and a Unification component (Hagoort 2013; Hagoort 2016), on a semantic memory containing knowledge about events and their typical participants (McRae and Matsuki 2009) and on a cost for the composition of unified event representations, showed the ability of modeling some complexity phenomena related to event knowledge.

In this conclusive section, we meditate on the contribution of the present work to the problem of complexity. Our approach, which brings to the table a new perspective focused on the semantic aspects related to event knowledge, will be compared to the other frameworks that we mentioned in the introduction. We then discuss possible future directions for complexity models, starting from their current limitations due to the lack of evaluation benchmarks and from the potential application of the new and powerful Deep Learning models.

4.1. A Semantic Perspective on Complexity

In the introductory chapter, we mentioned several accounts of linguistic complexity, situating our theoretical proposal in a trend of studies looking at complexity as a property of specific linguistic realizations, i.e. natural language sentences (Brunato 2015). Some important features in common with other models are the focus on the factors that make a particular sentence more or less difficult for human understanding and the evaluation with reference to psycholinguistic and neurolinguistic works on sentence processing.

Apart from these common points, the adopted perspective is different from, and somehow complementary to most of the previous work on the topic. Frameworks such as those proposed by Gibson 2000 and R. L. Lewis and Vasishth 2005 mainly look at syntactic factors, without taking into account semantics. The opposite can be said about our model, in which syntax is mostly used to identify aspects of the semantic structure of the event described by the sentence. However, in a constraint-based vision of processing in which the relevant information for each linguistic domain is represented separately (Blache 2016), it is conceivable that such models can be somehow integrated, and different complexity indexes could be used to account for different complexity sources.

The relationship with surprisal-based models is more complex. The relationship between surprisal and processing times has been described in the literature as a *causal bottleneck*: independently from the mental representations underlying language comprehension, processing difficulty predictions are affected only by the conditional probabilities that those representations determine (R. Levy 2008; Delogu, Crocker, and Drenhaus 2017).

In other words, surprisal-based models are somehow agnostic with respect to the content stored in the semantic memory: the only thing that matters are word probabilities.

Our model differs from such account, as the processing costs are considered to be, at least in part, construction-specific and not totally depending on conditional probabilities. One striking case where the two models lead to different prediction is logical metonymy, as the coercion operation is assumed to determine extra processing times: in a surprisal account, the extra costs are simply due to the low probability of the object noun, whereas construction-specific accounts generally explain them in terms of the recovery of a hidden event in order to repair a sort of semantic violation (a low thematic fit, in our framework).

In the experimental literature, the recent work by Delogu, Crocker, and Drenhaus 2017 proved that actually, when the predictability of the object noun is similar to the coerced one, the extra processing times can be largely accounted for by the surprisal (with coercion operation possibly affecting later processing stages). We are not aware of any experimental study bringing evidence that, in contrast, could help teasing apart surprisal and construction-specific costs, so it could well be the case that probabilities are the only things to be looked at.

On the other hand, in a more theoretical perspective, one aspect regarding surprisal-based models that is not totally clear is their ability to work with unattested data. There might be different reasons why a given linguistic expression is absent from a corpus: it could be rare, it might describe false facts, it could be too anomalous from a semantic point of view etc. A common criticism to probabilistic models coming from generative linguistics is, precisely, that one cannot distinguish between these cases on the

basis of word probabilities (Vecchi, Marelli, Zamparelli, and Baroni 2017).

In a computational perspective, a possible advantage for thematic fit models like ours is that they are, in theory, more robust to data sparsity, while probabilistic models might require complex smoothing techniques to deal with zero counts. Our results go in that direction, as thematic fit models did not have any coverage problem, while maintaining good levels of accuracy (e.g. see the experiments on the McElree and on the Traxler datasets).

However, it should be pointed out that our datasets were not the ideal ones for a comparison between the two types of models, as the probabilities had not been properly controlled, so that there could be significant differences between conditions. In the end, we cannot exclude that the higher processing costs for metonymies simply reflect a lower probability of the complement nouns, independently from the coercion operation.

Future experiments should aim at modeling datasets in which cloze probabilities are carefully controlled and similar across conditions, in order to properly address the issue.

4.2. Semantic Complexity Models: The Need for Robust Evaluation Benchmarks

Admittedly, our DSM for semantic complexity has been tested until now only on simple sentences from psycholinguistic experiments, which have been modeled as subject-verb-object triplets. The problem of semantic complexity in sentence processing was mainly treated in relation to the problem of argument typicality: typical argument combinations are easier to process, while for the unexpected ones it is more difficult to build a semantic representation (Baggio, Van Lambalgen, and Hagoort 2012). However, this equals to 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 sentential or to the general discourse context.

It should be pointed out that current approaches to the estimation of argument typicality also limit themselves to relatively easy tasks, and one of the main reasons is the scarcity of benchmark datasets. Studies focusing the automatic estimation of thematic fit have generally adopted an evaluation based on the correlation between system predictions and human ratings (Baroni and Lenci 2010; A. Sayeed and Demberg 2014; A. Sayeed, Demberg, and Shkadzko 2015; Greenberg, A. B. Sayeed, and Demberg 2015; Tilk, Demberg, A. Sayeed, Klakow, and Thater 2016; Santus, Chersoni, Lenci, and Blache 2017; see section 2.2.1), but most of these works did not take into account the dynamic aspect of the phenomenon, i.e. the fact that the plausibility of arguments changes as the other roles are filled. The gold standards datasets mostly consist of simple ratings of

verb-argument pairs in isolation, and do not take into account how the typicality scores change in function of the other event participants.

A few research papers (Lenci 2011, Tilk, Demberg, A. Sayeed, Klakow, and Thater 2016) paid attention to the aspect of the dynamic update of the expectations, but they were still strongly limited by the lack of strong benchmarks for the task. In the end, the binary classification on the Bicknell dataset turned out to be the only option for testing this aspect of the model. It is worth noticing that the Bicknell dataset alone might not be reliable for such a test: in our experimentations, a simple algorithm based on vector product and with no information on typical fillers achieved the top scores in the classification task (while it fell short on the other tasks on logical metonymy).

The recently-introduced DTFit dataset (Vassallo, Chersoni, Santus, Lenci, and Blache 2018) was meant to be a first step in the direction of defining a strong benchmark for thematic fit models. The dataset presents the following features:

— the inclusion of more complex argument combinations, such as *agent-patient-instrument*;

— the inclusion a wide variety of roles, such as *instruments, locations* and *time*;

— human ratings for the plausibility of each combination;

— annotations on the typicality of the combination (e.g. typical / non-typical).

Such features allow for a more satisfactory evaluation strategies, since typicality scores are referred to a complex event structure, and not just to single roles. Classification task are still possible, since all the dataset tuples

are labeled as either typical or non-typical, but they are not limited to agent-verb-patient triples. The first evaluations carried out on the dataset suggest that it is extremely difficult to achieve high correlation values with the ratings, especially on roles other than agent and patient.

Hopefully, future research will include also other sources of complexity beyond event-related argument typicality. For example, recent psycho- and neurolinguistic research proved that selectional restrictions violations lead to disruption patterns in eye-tracking data that are different from the ones caused by GEK violations (Warren, Milburn, Patson, and Dickey 2015), and that they cannot be modulated by the semantic relatedness with the sentential/discourse context (that is the case instead for the occurrence of simply atypical arguments: see Paczynski and Kuperberg 2012). Since the two violation types lead to different processing effects, ideally a model of complexity should be able to distinguish between them and to assign higher scores to the semantic violation determining the higher costs. The facilitation effect caused by the processing of semantically-related words in the general context should be taken into account as well, as it seems to be based on a type of semantic knowledge that differs from the two mentioned before (Federmeier and Kutas 1999; Metusalem, Kutas, Urbach, Hare, McRae, and Elman 2012; Paczynski and Kuperberg 2012). Finally, at the best of our knowledge, the role played by modifiers has not been explored yet in the experimental literature on sentence processing, but it is presumable that they are also involved in typicality relations and in the generation of expectations on the upcoming input.

Ideally, future systems aiming at modeling semantic complexity should be able to deal with all these factors, and new benchmarks will have to be built with this objective in mind.

4.3. Embeddings and Deep Learning: What Contribution to Complexity Modeling?

From the point of view of the algorithm, the core of our current complexity model is a traditional DSM, since at the beginning of our work they presented some important advantages over the recently-born word embeddings. First, we aimed at having a theoretical and computational framework in which the constructions stored in the semantic memory (i.e. in the GEK) could have an explicit and interpretable representation, and the dimensions of dense word embeddings are not been considered interpretable for a long time. Secondly, our architecture heavily relies on a thematic fit estimation component, one of the few tasks in which the embeddings seemed to be lagging behind traditional vector spaces (Baroni, Dinu, and Kruszewski 2014).

However, the advances brought by the Deep Learning revolution in the last few years cannot be ignored (C. D. Manning 2015) and they could quickly change this scenario. A problem with traditional DSMs is that it takes a lot of time to train the models and it is difficult to find the optimal parameters, and this is especially true for our framework, where we have to extract event representations for populating the GEK and to estimate thematic fit scores for several argument combinations. One advantage of word embeddings is that, for many of the available software packages, training parameters have already been set to optimal values by the algorithm engineers (O. Levy, Y. Goldberg, and Dagan 2015), and this makes the whole training process -from raw text to dense vector representations- more smooth and compact. Also the computation of argument filler plausibility could greatly benefit from neural network-based supervised training

(Tilk, Demberg, A. Sayeed, Klakow, and Thater 2016).

Concerning the thematic fit task, there are at least two novelties in recent NLP research that need to be mentioned. The first concerns the extension of the embedding models to structured contexts and their increasing popularity. The original comparisons between count-based models and the word embeddings were based on the original architecture by Mikolov and colleagues, which was trained on linear, bag-of-words context windows (Mikolov, Chen, G. Corrado, and Jeffrey Dean 2013; Mikolov, Sutskever, Chen, G. S. Corrado, and Jeff Dean 2013), with the result of a mediocre performance of the embeddings on thematic fit estimation. But then research work like O. Levy and Y. Goldberg 2014a showed that, when embeddings are trained on structured contexts (e.g. syntactic dependencies), they model exactly the functional type of similarity that we look for while comparing role prototypes and candidate fillers in the our task. [1] Although, at the best of our knowledge, systematic comparisons between traditional structured DSMs (e.g. Distributional Memory) and dependency-based embeddings have not been carried out yet, it is reasonable to think that structure would help embeddings to improve their performance in the task. [2]

The second important novelty consists in the recent attempts of modeling event participants with neural networks (Tilk, Demberg, A. Sayeed, Klakow, and Thater 2016; Weber, Balasubramanian, and Chambers 2018). [3]

1. Recall the distinction, already in Turney 2012, between i) the similarity between words that tend to occur together as they are both related to the same topic (*domain similarity*) and ii) the similarity between words that tend to carry out the same carry out the same functions in relation to other words (*function similarity*). See also Section 2.1.4. for reference.

2. It is worth mentioning that a recent study by Heinzerling, Moosavi, and Strube 2017 made use of dependency-based embeddings to incorporate selectional preference information in a coreference resolution system. Although the comparison between count-based models and predictive models was out of the scope of their contribution, their results are a proof that embeddings trained on structured contexts can be efficiently used for modeling selectional preferences.

3. Another related trend of research work, which is worth mentioning here, makes use of dis-

The work by Tilk, Demberg, A. Sayeed, Klakow, and Thater 2016 showed some impressive results: the system, trained on a corpus labeled with semantic roles, outperformed all the previous methods in almost all the traditional thematic fit datasets by a significant margin. There is without any doubt is an impressive proof of the potential of Deep Learning for tasks related to event knowledge modeling. However, there is still room for improvement, in the sense that the same model perform similarly to Lenci's system (Lenci 2011) in the task of the update of argument expectations. Moreover, the benchmark available to Tilk and colleagues (and the same goes for the following work by Weber, Balasubramanian, and Chambers 2018, who also tested their system in a similar task) were still limited to subject-verb-object triples, as in the Bicknell dataset.

Again, future developments of such models will have to deal with test datasets providing examples of more complex and challenging event-participants interactions.

tributed vector representations and neural networks for modeling script knowledge and predicting upcoming events in narrative chains (Granroth-Wilding and Clark 2016; Pichotta and Mooney 2016; Z. Wang, Y. Zhang, and Chang 2017).

4.4. Conclusive Remarks

In this thesis, we have introduced the topic of semantic complexity in language processing, and we have discussed it in the light of some important research lines in modern experimental linguistics, namely the neurobiological architecture of Memory, Unification and Control, proposed by Hagoort and colleagues (Baggio and Hagoort 2011; Baggio, Van Lambalgen, and Hagoort 2012; Hagoort 2013; Hagoort 2016); the psycholinguistic studies on the role of the Generalized Event Knowledge (GEK) in human sentence processing (McRae, M. Spivey-Knowlton, and Tanenhaus 1998; McRae, Hare, Elman, and T. Ferretti 2005; Hare, M. Jones, Thomson, Kelly, and McRae 2009; Bicknell, Elman, Hare, McRae, and Kutas 2010; Matsuki, Chow, Hare, Elman, Scheepers, and McRae 2011; Metusalem, Kutas, Urbach, Hare, McRae, and Elman 2012); Jeff Elman's words-as-cues hypothesis (Elman 2009; Elman 2014) and the hierarchical generative frameworks of language processing (Kleinschmidt and Jaeger 2015; M. Brown and Kuperberg 2015; Kuperberg 2016), which share the idea of the comprehension process as a cue-to-category mapping problem. The framework presented here borrows ideas from all these sources of inspiration, and proposes an integration of a theoretical account of semantic complexity with the implementation of a concrete computational model.

While the distributional model heavily relies on the component of thematic fit estimation, which had been already proposed for modeling the update of expectations in sentence processing (Lenci 2011), an important element of novelty consists in the introduction of a simple GEK component, based on the storage and retrieval of ready-made event representations. Events are modeled by means of the recently-introduced syntactic

joint contexts (Chersoni, Santus, Lenci, Blache, and C.-R. Huang 2016), that correspond to a surface approximation of an event structure, directly extracted from parsed corpora.

We proposed an evaluation by means of a comparison with the results of studies in sentence processing, in order to link the outputs of our model with the experimental data. Among our results, one of our complexity models was the first to simultaneously deal with the two aspects of the phenomenon of logical metonymy (the increased processing costs caused by coercion, and the retrieval of a covert event).

Future work on our model will aim at the application to other semantic tasks involving event knowledge, such as the detection of different type of anomalies (e.g. violations of the GEK vs. violations of selectional restrictions), the recovery of implicit arguments and of bridging inferences. Going in this direction, mandatory steps will presumably be the creation of new and more realistic datasets, on the one hand, and the exploitation of the last generation of Deep Learning models for NLP, on the other hand.

Bibliography

[Alm04] Roberto G. de Almeida. "The Effect of Context on the Processing of Type-shifting Verbs". In: *Brain and Language* 90 (2004), pp. 249–261 (cit. on p. 106).

[AD08] Roberto G. de Almeida and Veena D. Dwivedi. "Coercion without Lexical Decomposition: Type-shifting Effects Revisited". In: *Canadian Journal of Linguistics* 53.2/3 (2008), pp. 301–326 (cit. on p. 106).

[Ash15] Nicholas Asher. "Types, Meanings and Coercions in Lexical Semantics". In: *Lingua* 157 (2015), pp. 66–82 (cit. on p. 104).

[AJ17] Fatemeh Torabi Asr and Michael Jones. "An Artificial Language Evaluation of Distributional Semantic Models". In: *Proceedings of CONLL*. 2017 (cit. on p. 57).

[AWJ16] Fatemeh Torabi Asr, Jon A Willits, and Michael Jones. "Comparing Predictive and Co-Occurrence Based Models of Lexical Semantics Trained on Child-Directed Speech". In: *Proceedings of CogSci*. 2016 (cit. on p. 57).

[BH11] Giosuè Baggio and Peter Hagoort. "The Balance between Memory and Unification in Semantics: A Dynamic Account of the N400". In: *Language and Cognitive Processes* 26.9 (2011), pp. 1338–1367 (cit. on pp. 19, 42, 92, 131).

[BVH12] Giosuè Baggio, Michiel Van Lambalgen, and Peter Hagoort. "The Processing Consequences of Compositionality". In: *The*

Oxford Handbook of Compositionality. Oxford (2012), pp. 657–674 (cit. on pp. 19, 20, 73, 125, 131).

[BBZ14] Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. "Frege in Space: A Program of Compositional Distributional Semantics". In: *LiLT (Linguistic Issues in Language Technology)* 9 (2014) (cit. on p. 68).

[Bar+09] Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora". In: *Language Resources and Evaluation* 43.3 (2009), pp. 209–226 (cit. on p. 92).

[BDK14] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. "Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors." In: *Proceedings of ACL*. 2014, pp. 238–247 (cit. on pp. 50, 53, 54, 57, 62, 128).

[BL10] Marco Baroni and Alessandro Lenci. "Distributional Memory: A General Framework for Corpus-Based Semantics". In: *Computational Linguistics* 36.4 (2010), pp. 673–721 (cit. on pp. 52, 58, 60, 63, 64, 85, 93, 97, 125).

[BZ10] Marco Baroni and Roberto Zamparelli. "Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space". In: *Proceedings of EMNLP*. 2010 (cit. on p. 68).

[Ber+13] Raffaella Bernardi, Georgiana Dinu, Marco Marelli, and Marco Baroni. "A Relatedness Benchmark to Test the Role of Deter-

miners in Compositional Distributional Semantics." In: *Proceedings of ACL*. 2013 (cit. on p. 69).

[Bic+10] Klinton Bicknell, Jeffrey L Elman, Mary Hare, Ken McRae, and Marta Kutas. "Effects of Event Knowledge in Processing Verbal Arguments". In: *Journal of memory and language* 63.4 (2010), pp. 489–505 (cit. on pp. 29, 30, 32, 43, 63, 86, 94, 95, 131).

[Bla11] Philippe Blache. "Evaluating Language Complexity in Context: New Parameters for a Constraint-Based Model". In: *CSLP-11, Workshop on Constraint Solving and Language Processing*. 2011 (cit. on pp. 10, 11).

[Bla16] Philippe Blache. "Representing Syntax by Means of Properties: A Formal Framework for Descriptive Approaches". In: *Journal of Language Modelling* 4.2 (2016), pp. 183–224 (cit. on p. 122).

[BL12a] William Blacoe and Mirella Lapata. "A Comparison of Vector-Based Representations for Semantic Composition". In: *Proceedings of EMNLP-CONLL*. 2012 (cit. on p. 67).

[Boj+16] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching Word Vectors with Subword Information". In: *arXiv preprint arXiv:1607.04606* (2016) (cit. on p. 57).

[BVA15] Antoine Bride, Tim Van de Cruys, and Nicholas Asher. "A Generalisation of Lexical Functions for Composition in Distributional Semantics." In: *Proceedings of ACL*. 2015 (cit. on p. 69).

[BK15] Meredith Brown and Gina R Kuperberg. “A Hierarchical Generative Framework of Language Processing: Linking Language Perception, Interpretation, and Production Abnormalities in Schizophrenia”. In: *Frontiers in Human Neuroscience* 9 (2015) (cit. on pp. 33, 40, 131).

[Bru15] Dominique Pierina Brunato. “A Study on Linguistic Complexity from a Computational Linguistics Perspective. A Corpus-Based Investigation of Italian Bureaucratic Texts”. PhD thesis. University of Siena, 2015 (cit. on pp. 10, 122).

[BL12b] John Bullinaria and Joseph Levy. “Extracting Semantic Representations from Word Co-Occurrence Statistics: Stop-Lists, Stemming, and SVD”. In: *Behavior Research Methods* 44.3 (2012), pp. 890–907 (cit. on pp. 51, 53).

[BL97] Curt Burgess and Kevin Lund. “Representing Abstract Words and Emotional Connotation in a High-Dimensional Memory Space”. In: *Proceedings of the Cognitive Science Society*. 1997, pp. 61–66 (cit. on p. 61).

[Byb13] Joan L Bybee. “Usage-Based Theory and Exemplar Representations of Constructions”. In: *The Oxford Handbook of Construction Grammar*. 2013 (cit. on p. 21).

[CPN15] José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. “NASARI: a Novel Approach to a Semantically-Aware Representation of Items.” In: *Proceedings of HLT-NAACL*. 2015, pp. 567–577 (cit. on p. 50).

[Car02] Robyn Carston. *Thoughts and Utterances*. Oxford: Blackwell, 2002 (cit. on p. 106).

[CBL16] Emmanuele Chersoni, Philippe Blache, and Alessandro Lenci. "Towards a Distributional Model of Semantic Complexity". In: *Proceedings of the COLING Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*. 2016, pp. 12–22 (cit. on pp. 75, 86).

[CLB17] Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. "Logical Metonymy in a Distributional Model of Sentence Comprehension". In: *Proceedings of *SEM*. 2017, pp. 168–177 (cit. on pp. 75, 86).

[Che+17] Emmanuele Chersoni, Enrico Santus, Philippe Blache, and Alessandro Lenci. "Is Structure Necessary for Modeling Argument Expectations in Distributional Semantics?" In: *Proceedings of IWCS* (2017) (cit. on pp. 86, 88, 97, 102).

[Che+16] Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. "Representing Verbs with Rich Contexts: An Evaluation on Verb Similarity". In: *Proceedings of EMNLP*. 2016 (cit. on pp. 78, 92, 132).

[Cho57] Noam Chomsky. "Syntactic Structures". In: *Language* (1957) (cit. on p. 17).

[CH90] Kenneth Ward Church and Patrick Hanks. "Word Association Norms, Mutual Information, and Lexicography". In: *Computational linguistics* 16.1 (1990), pp. 22–29 (cit. on p. 51).

[CP07] Stephen Clark and Stephen Pulman. "Combining Symbolic and Distributional Models of Meaning". In: *Proceedings of the AAAI Spring Symposium on Quantum Interaction, Stanford, CA, 2007*. 2007 (cit. on p. 66).

[CS08] Charles Clifton and Adrian Staub. “Parallelism and Competition in Syntactic Ambiguity Resolution”. In: *Language and Linguistics Compass* 2.2 (2008), pp. 234–250 (cit. on p. 23).

[Col+11] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. “Natural Language Processing (Almost) from Scratch”. In: *Journal of Machine Learning Research* 12.Aug (2011), pp. 2493–2537 (cit. on p. 64).

[CJ05] Peter W Culicover and Ray Jackendoff. *Simpler Syntax*. Oxford University Press, 2005 (cit. on p. 22).

[Dee+90] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. “Indexing by Latent Semantic Analysis”. In: *Journal of the American society for information science* 41.6 (1990), p. 391 (cit. on p. 52).

[DCD17] Francesca Delogu, Matthew W Crocker, and Heiner Drenhaus. “Teasing Apart Coercion and Surprisal: Evidence from Eye-Movements and ERPs”. In: *Cognition* 161 (2017), pp. 46–59 (cit. on pp. 111, 122, 123).

[DK08] Vera Demberg and Frank Keller. “Data from Eye-Tracking Corpora as Evidence for Theories of Syntactic Processing Complexity”. In: *Cognition* 109.2 (2008), pp. 193–210 (cit. on pp. 13, 15).

[DPB13] Georgiana Dinu, Nghia The Pham, and Marco Baroni. “General Estimation and Evaluation of Compositional Distributional Semantic Models”. In: *Proceedings of the ACL Workshop on Continuous Vector Space Models and their Compositionality*. 2013 (cit. on p. 69).

[Elm09] Jeffrey L Elman. "On the Meaning of Words and Dinosaur Bones: Lexical Knowledge without a Lexicon". In: *Cognitive Science* 33.4 (2009), pp. 547–582 (cit. on pp. 33, 34, 37, 43, 72, 106, 131).

[Elm14] Jeffrey L Elman. "Systematicity in the Lexicon: On Having your Cake and Eating It Too". In: *The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*. Ed. by Paco Calvo and John Symons. Cambridge, MA: The MIT Press, 2014 (cit. on pp. 33, 34, 36, 43, 106, 131).

[Erk07] Katrin Erk. "A Simple, Similarity-Based Model for Selectional Preferences". In: *Proceedings of ACL*. 2007 (cit. on p. 63).

[EP08] Katrin Erk and Sebastian Padó. "A Structured Vector Space Model for Word Meaning in Context". In: *Proceedings of EMNLP*. 2008 (cit. on p. 67).

[EPP10] Katrin Erk, Sebastian Padó, and Ulrike Padó. "A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences". In: *Computational Linguistics* 36.4 (2010), pp. 723–763 (cit. on pp. 63, 85).

[Ett+16] Allyson Ettinger, Naomi H Feldman, Philip Resnik, and Colin Phillips. "Modeling N400 Amplitude Using Vector Space Models of Word Representation". In: *Proceedings of CogSci*. 2016, pp. 1445–1450 (cit. on p. 62).

[EL16] Allyson Ettinger and Tal Linzen. "Evaluating Vector Space Models Using Human Semantic Priming Results". In: *Proceedings of the Workshop on Evaluating Vector Space Representations for NLP*. 2016, pp. 72–77 (cit. on p. 61).

[Eve04] Stefan Evert. “The Statistics of Word Cooccurrences: Word Pairs and Collocations”. PhD thesis. 2004 (cit. on pp. 52, 93).

[Fal11] Ingrid L. Falkum. “A Pragmatic Account of ‘Logical Metonymy’”. In: *Proceedings of Metonymy 2011*. Stuttgart, Germany, 2011, pp. 11–17 (cit. on pp. 105, 106).

[FK99] Kara D Federmeier and Marta Kutas. “A Rose by any Other Name: Long-Term Memory Structure and Sentence Processing”. In: *Journal of memory and Language* 41.4 (1999), pp. 469–495 (cit. on pp. 19, 127).

[FC86] Fernanda Ferreira and Charles Clifton. “The Independence of Syntactic Processing”. In: *Journal of memory and language* 25.3 (1986), pp. 348–368 (cit. on p. 23).

[FMH01] Todd R Ferretti, Ken McRae, and Andrea Hatherell. “Integrating Verbs, Situation Schemas, and Thematic Role Concepts”. In: *Journal of Memory and Language* 44.4 (2001), pp. 516–547 (cit. on pp. 27, 34, 65).

[FL99] Jerry Fodor and Ernie Lepore. “All at Sea in Semantic Space: Churchland on Meaning Similarity”. In: *The Journal of Philosophy* 96.8 (1999), pp. 381–403 (cit. on p. 62).

[Fra87] Lyn Frazier. “Sentence Processing: A Tutorial Review.” In: (1987) (cit. on pp. 22, 23).

[FM08] Steven Frisson and Brian McElree. “Complement Coercion is not Modulated by Competition: Evidence from Eye Movements.” In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34.1 (2008), pp. 1–11 (cit. on p. 105).

[FMG15] Richard Futrell, Kyle Mahowald, and Edward Gibson. "Large-Scale Evidence of Dependency Length Minimization in 37 Languages". In: *Proceedings of the National Academy of Sciences* 112.33 (2015), pp. 10336–10341 (cit. on p. 13).

[Gib98] Edward Gibson. "Linguistic Complexity: Locality of Syntactic Dependencies". In: *Cognition* 68.1 (1998), pp. 1–76 (cit. on p. 12).

[Gib00] Edward Gibson. "The Dependency Locality Theory: A Distance-Based Theory of Linguistic Complexity". In: *Image, language, brain* (2000), pp. 95–126 (cit. on pp. 12, 122).

[GT10] Daniel Gildea and David Temperley. "Do Grammars Minimize Dependency Length?" In: *Cognitive Science* 34.2 (2010), pp. 286–310 (cit. on p. 13).

[Gol95] Adele E Goldberg. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, 1995 (cit. on p. 21).

[Gol06] Adele E Goldberg. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press on Demand, 2006 (cit. on p. 21).

[GC16] Mark Granroth-Wilding and Stephen Clark. "What Happens Next? Event Prediction Using a Compositional Neural Network Model". In: *Proceedings of AAAI*. 2016 (cit. on p. 130).

[GDS15] Clayton Greenberg, Vera Demberg, and Asad Sayeed. "Verb Polysemy and Frequency Effects in Thematic Fit Modeling". In: *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*. 2015 (cit. on p. 65).

[GSD15] Clayton Greenberg, Asad B Sayeed, and Vera Demberg. "Improving Unsupervised Vector-Space Thematic Fit Evaluation via Role-Filler Prototype Clustering." In: *Proceedings of HLT-NAACL*. 2015 (cit. on pp. 65, 125).

[Gre+13] Edward Grefenstette, Georgiana Dinu, Yao-Zhong Zhang, Mehrnoosh Sadrzadeh, and Marco Baroni. "Multi-Step Regression Learning for Compositional Distributional Semantics". In: *Proceedings of IWCS*. 2013 (cit. on p. 68).

[Gre94] Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, 1994 (cit. on p. 58).

[Gue10] Emiliano Guevara. "A Regression Model of Adjective-Noun Compositionality in Distributional Semantics". In: *Proceedings of the ACL Workshop on GEometrical Models of Natural Language Semantics*. 2010 (cit. on p. 68).

[Gup+15] Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. "Distributional Vectors Encode Referential Attributes." In: *Proceedings of EMNLP*. 2015, pp. 12–21 (cit. on p. 62).

[GBP17] Abhijeet Gupta, Gemma Boleda, and Sebastian Padó. "Distributed Prediction of Relations for Entities: The Easy, The Difficult, and The Impossible". In: *Proceedings of *SEM 2017*. 2017, pp. 104–109 (cit. on p. 62).

[GUP15] Abhijeet Gupta, Jason Utt, and Sebastian Padó. "Dissecting the Practical Lexical Function Model for Compositional Distributional Semantics". In: *Proceedings of Starsem*. 2015, pp. 153–158 (cit. on p. 70).

[Hag13] Peter Hagoort. “MUC (Memory, Unification, Control) and Beyond”. In: *Frontiers in Psychology* 4 (2013), p. 416 (cit. on pp. 21, 73, 77, 121, 131).

[Hag16] Peter Hagoort. “MUC (Memory, Unification, Control): A Model on the Neurobiology of Language Beyond Single Word Processing”. In: *Neurobiology of language*. Elsever, 2016, pp. 339–347 (cit. on pp. 21, 22, 73, 77, 121, 131).

[Hal01] John Hale. “A Probabilistic Earley Parser as a Psycholinguistic Model”. In: *Proceedings of NAACL-HLT*. Association for Computational Linguistics. 2001, pp. 1–8 (cit. on p. 13).

[Hal16] John Hale. “Information-Theoretical Complexity Metrics”. In: *Language and Linguistics Compass* 10.9 (2016), pp. 397–412 (cit. on pp. 13, 14).

[Har+09] Mary Hare, Michael Jones, Caroline Thomson, Sarah Kelly, and Ken McRae. “Activating Event Knowledge”. In: *Cognition* 111.2 (2009), pp. 151–167 (cit. on pp. 27, 28, 131).

[Har68] Zellig S Harris. “Mathematical Structures of Language”. In: (1968) (cit. on p. 47).

[Har70] Zellig S Harris. “Distributional Structure”. In: *Papers in Structural and Transformational Linguistics* (1970), pp. 775–794 (cit. on p. 47).

[HMS17] Benjamin Heinzerling, Nafise Sadat Moosavi, and Michael Strube. “Revisiting Selectional Preferences for Coreference Resolution”. In: 2017 (cit. on p. 129).

[HRK16] Felix Hill, Roi Reichart, and Anna Korhonen. "Simlex-999: Evaluating semantic models with (genuine) similarity estimation". In: *Computational Linguistics* (2016) (cit. on p. 59).

[Jac97] Ray Jackendoff. *The Architecture of the Language Faculty*. 28. MIT Press, 1997 (cit. on pp. 22, 71, 104, 105).

[Jac03] Ray Jackendoff. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. 2003 (cit. on pp. 22, 24).

[JJ15] Brendan T Johns and Michael N Jones. "Generating Structure from Experience: A Retrieval-Based Model of Language Processing." In: *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale* 69.3 (2015), p. 233 (cit. on p. 39).

[JKM06] Michael N Jones, Walter Kintsch, and Douglas JK Mewhort. "High-Dimensional Semantic Space Accounts of Priming". In: *Journal of memory and language* 55.4 (2006), pp. 534–552 (cit. on p. 61).

[JM14] Dan Jurafsky and James H Martin. *Speech and language processing*. Vol. 3. Pearson London: 2014 (cit. on p. 99).

[Kaa07] Edith Kaan. "Event-Related Potentials and Language Processing: A Brief Overview". In: *Language and Linguistics Compass* 1.6 (2007), pp. 571–591 (cit. on pp. 18, 19).

[Kam13] Hans Kamp. *Meaning and the Dynamics of Interpretation: Selected Papers by Hans Kamp*. Leiden-Boston: Brill, 2013 (cit. on p. 81).

[Kin88] Walter Kintsch. “The Role of Knowledge in Discourse Comprehension: A Construction-Integration Model.” In: *Psychological Review* 95.2 (1988), p. 163 (cit. on p. 38).

[KJ15] Dave F Kleinschmidt and T Florian Jaeger. “Robust Speech Perception: Recognize the Familiar, Generalize to the Similar, and Adapt to the Novel.” In: *Psychological Review* 122.2 (2015), p. 148 (cit. on pp. 41, 43, 131).

[KH09] Kristina Küper and Martin Heil. “Electrophysiology Reveals Semantic Priming at a Short SOA Irrespective of Depth of Prime Processing”. In: *Neuroscience Letters* 453.2 (2009), pp. 107–111 (cit. on p. 27).

[Kup16] Gina R Kuperberg. “Separate Streams or Probabilistic Inference? What the N400 Can Tell Us About the Comprehension of Events”. In: *Language, Cognition and Neuroscience* 31.5 (2016), pp. 602–616 (cit. on pp. 40, 43, 82, 131).

[KJ16] Gina R Kuperberg and T Florian Jaeger. “What Do We Mean by Prediction in Language Comprehension?” In: *Language, Cognition and Neuroscience* 31.1 (2016), pp. 32–59 (cit. on pp. 40, 43, 76).

[KF00] Marta Kutas and Kara D Federmeier. “Electrophysiology Reveals Semantic Memory Use in Language Comprehension”. In: *Trends in cognitive sciences* 4.12 (2000), pp. 463–470 (cit. on pp. 19, 20).

[LD97] Thomas K Landauer and Susan T Dumais. “A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of Ac-

quisition, Induction, and Representation of Knowledge." In: *Psychological review* 104.2 (1997), p. 211 (cit. on pp. 52, 61).

[LKS03] Mirella Lapata, Frank Keller, and Christoph Scheepers. "Intra-Sentential Context Effects on the Interpretation of Logical Metonymy". In: *Cognitive Science* 27.4 (2003), pp. 649–668 (cit. on p. 107).

[LL03] Mirella Lapata and Alex Lascarides. "A Probabilistic Account of Logical Metonymy". In: *Computational Linguistics* 29.2 (2003), pp. 261–315 (cit. on p. 107).

[LE17] Gabriella Lapesa and Stefan Evert. "Large-Scale Evaluation of Dependency-Based DSMs: Are They Worth the Effort?" In: *Proceedings of EACL* (2017) (cit. on p. 60).

[LC98] Alex Lascarides and Ann Copestake. "Pragmatics and Word Meaning". In: *Journal of Linguistics* 34 (1998), pp. 378–414 (cit. on p. 105).

[LL17] Gianluca E. Lebani and Alessandro Lenci. "A Distributional Model of Verb-Specific Semantic Role Inferences". In: *Language, Cognition and Computational Models*. Ed. by Thierry Poibeau and Aline Villavicencio. Cambridge University Press, 2017 (cit. on pp. 29, 61).

[Lee92] Geoffrey Neil Leech. "100 million words of English: the British National Corpus (BNC)". In: (1992) (cit. on p. 92).

[Len08] Alessandro Lenci. "Distributional Semantics in Linguistic and Cognitive Research". In: *Italian Journal of Linguistics* 20.1 (2008), pp. 1–31 (cit. on pp. 11, 46, 61).

[Len11] Alessandro Lenci. "Composing and Updating Verb Argument Expectations: A Distributional Semantic Model". In: *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics. 2011, pp. 58–66 (cit. on pp. 29, 63, 65, 86, 93–95, 100, 102, 126, 130, 131).

[LG14a] Omer Levy and Yoav Goldberg. "Dependency-Based Word Embeddings." In: *Proceedings of ACL*. 2014, pp. 302–308 (cit. on pp. 54, 59, 129).

[LG14b] Omer Levy and Yoav Goldberg. "Neural Word Embedding as Implicit Matrix Factorization". In: *Advances in Neural Information Processing Systems*. 2014, pp. 2177–2185 (cit. on p. 56).

[LGD15] Omer Levy, Yoav Goldberg, and Ido Dagan. "Improving Distributional Similarity with Lessons Learned from Word Embeddings". In: *Transactions of ACL* 3 (2015), pp. 211–225 (cit. on pp. 51, 53, 54, 57, 128).

[Lev08] Roger Levy. "Expectation-Based Syntactic Comprehension". In: *Cognition* 106.3 (2008), pp. 1126–1177 (cit. on pp. 13, 122).

[Lew+04] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. "Rcv1: A New Benchmark Collection for Text Categorization Research". In: *Journal of machine learning research* 5.Apr (2004), pp. 361–397 (cit. on p. 92).

[LV05] Richard L Lewis and Shravan Vasishth. "An Activation-Based Model of Sentence Processing as Skilled Memory Retrieval".

In: *Cognitive science* 29.3 (2005), pp. 375–419 (cit. on pp. 13, 122).

[Lin98] Dekang Lin. "Automatic Retrieval and Clustering of Similar Words". In: *Proceedings of COLING*. 1998, pp. 768–774 (cit. on p. 58).

[Man15] Christopher D Manning. "Computational Linguistics and Deep Learning". In: *Computational Linguistics* 41.4 (2015), pp. 701–707 (cit. on p. 128).

[Mat+11] Kazunaga Matsuki, Tracy Chow, Mary Hare, Jeffrey L Elman, Christoph Scheepers, and Ken McRae. "Event-Based Plausibility Immediately Influences On-Line Language Comprehension." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37.4 (2011), p. 913 (cit. on pp. 30, 32, 34, 43, 131).

[McE+01] Brian McElree, Matthew J Traxler, Martin J Pickering, Rachel E Seely, and Ray Jackendoff. "Reading Time Evidence for Enriched Composition". In: *Cognition* 78 (2001), B17–B25 (cit. on pp. 104, 107–109, 113, 119).

[McR+05] Ken McRae, Mary Hare, Jeffrey L Elman, and Todd Ferretti. "A Basis for Generating Expectancies for Verbs from Nouns". In: *Memory & Cognition* 33.7 (2005), pp. 1174–1184 (cit. on pp. 28, 131).

[MM09] Ken McRae and Kazunaga Matsuki. "People Use their Knowledge of Common Events to Understand Language, and Do So as Quickly as Possible". In: *Language and linguistics compass*

3.6 (2009), pp. 1417–1429 (cit. on pp. 21, 24, 28, 32, 43, 74, 121).

[MM13] Ken McRae and Kazunaga Matsuki. "Constraint-Based Models of Sentence Processing". In: *Sentence processing* 519 (2013), pp. 51–77 (cit. on p. 25).

[MST98] Ken McRae, Michael Spivey-Knowlton, and Michael Tanenhaus. "Modeling the Influence of Thematic Fit (and other Constraints) in On-Line Sentence Comprehension". In: *Journal of Memory and Language* 38.3 (1998), pp. 283–312 (cit. on pp. 25, 26, 29, 63, 72, 84, 131).

[McW01] John McWhorter. "The World's Simplest Grammars Are Creole Grammars". In: *Linguistic Typology* 5.2 (2001), pp. 125–66 (cit. on p. 10).

[MAM84] Douglas L Medin, Mark W Altom, and Timothy D Murphy. "Given versus Induced Category Representations: Use of Prototype and Exemplar Information in Classification." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10.3 (1984), p. 333 (cit. on p. 42).

[Met+12] Ross Metusalem, Marta Kutas, Thomas P Urbach, Mary Hare, Ken McRae, and Jeffrey L Elman. "Generalized Event Knowledge Activation During Online Sentence Comprehension". In: *Journal of memory and language* 66.4 (2012), pp. 545–567 (cit. on pp. 30–32, 34, 127, 131).

[Mik+13a] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space".

In: *arXiv preprint arXiv:1301.3781* (2013) (cit. on pp. 54, 55, 129).

[Mik+13b] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems*. 2013, pp. 3111–3119 (cit. on pp. 54, 129).

[MC91] George A Miller and Walter G Charles. "Contextual Correlates of Semantic Similarity". In: *Language and Cognitive Processes* 6.1 (1991), pp. 1–28 (cit. on p. 61).

[ML08] Jeff Mitchell and Mirella Lapata. "Vector-Based Models of Semantic Composition." In: *Proceedings of ACL*. 2008 (cit. on p. 67).

[ML09] Jeff Mitchell and Mirella Lapata. "Language Models Based on Semantic Composition". In: *Proceedings of EMNLP*. 2009 (cit. on p. 67).

[ML10] Jeff Mitchell and Mirella Lapata. "Composition in Distributional Models of Semantics". In: *Cognitive Science* 34.8 (2010), pp. 1388–1429 (cit. on pp. 63, 66, 68, 98, 99, 101).

[Mon74] Richard Montague. "English as a Formal Language". In: *Formal Philosophy: Selected Papers of Richard Montague*. Ed. by Richmond H. Thomason. Yale University Press, 1974, pp. 188–222 (cit. on p. 68).

[Nee91] James H Neely. "Semantic Priming Effects in Visual Word Recognition: A Selective Review of Current Findings and The-

ories". In: *Basic Processes in Reading: Visual Word Recognition* 11 (1991), pp. 264–336 (cit. on p. 27).

[NV06] Mante S Nieuwland and Jos JA Van Berkum. "When Peanuts Fall in Love: N400 Evidence for the Power of Discourse". In: *Journal of cognitive neuroscience* 18.7 (2006), pp. 1098–1111 (cit. on p. 24).

[PK11] Martin Paczynski and Gina R Kuperberg. "Electrophysiological Evidence for Use of the Animacy Hierarchy, but not Thematic Role Assignment, During Verb-Argument Processing". In: *Language and Cognitive Processes* 26.9 (2011), pp. 1402–1456 (cit. on p. 42).

[PK12] Martin Paczynski and Gina R Kuperberg. "Multiple Influences of Semantic Memory on Sentence Processing: Distinct Effects of Semantic Relatedness on Violations of Real-World Event/State Knowledge and Animacy Selection Restrictions". In: *Journal of memory and language* 67.4 (2012), pp. 426–448 (cit. on pp. 42, 127).

[PL07] Sebastian Padó and Mirella Lapata. "Dependency-Based Construction of Semantic Space Models". In: *Computational Linguistics* 33.2 (2007), pp. 161–199 (cit. on p. 58).

[Pad07] Ulrike Padó. "The Integration of Syntax and Semantic Plausibility in a Wide-Coverage Model of Human Sentence Processing". PhD thesis. 2007 (cit. on p. 63).

[PL02] Patrick Pantel and Dekang Lin. "Discovering Word Senses from Text". In: *Proceedings of the ACM SIGKDD International*

Conference on Knowledge discovery and data mining. ACM. 2002, pp. 613–619 (cit. on p. 51).

[PPB14] Denis Paperno, Nghia The Pham, and Marco Baroni. "A Practical and Linguistically-Motivated Approach to Compositional Distributional Semantics". In: *Proceedings of ACL*. 2014 (cit. on p. 69).

[Par08] Mikael Parkvall. "The Simplicity of Creoles in a Cross-Linguistic Perspective". In: *Language complexity. Typology, contact, change* (2008), pp. 265–285 (cit. on p. 10).

[Pel94] Francis Jeffry Pelletier. "The Principle of Semantic Compositionality". In: *Topoi* 13.1 (1994), pp. 11–24 (cit. on p. 17).

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher Manning. "Glove: Global Vectors for Word Representation". In: *Proceedings of EMNLP*. 2014, pp. 1532–1543 (cit. on p. 57).

[PM16] Karl Pichotta and Raymond J Mooney. "Learning Statistical Scripts with LSTM Recurrent Neural Networks." In: *Proceedings of AAAI*. 2016 (cit. on p. 130).

[PJN13] Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. "Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity." In: *Proceedings of ACL*. 2013, pp. 1341–1351 (cit. on p. 50).

[Pus95] James Pustejovsky. *The Generative Lexicon*. Cambridge, MA: The MIT Press, 1995 (cit. on p. 105).

[PNK08] Joel Pynte, Boris New, and Alan Kennedy. "On-Line Contextual Influences During Reading Normal Text: A Multiple-Regression Analysis". In: *Vision Research* 48.21 (2008), pp. 2172–2183 (cit. on p. 15).

[QJA12] Ting Qian, T Florian Jaeger, and Richard N Aslin. "Learning to Represent a Multi-Context Environment: More than Detecting Changes". In: *Frontiers in Psychology* 3 (2012) (cit. on p. 38).

[RM14] Milena Rabovsky and Ken McRae. "Simulating the N400 ERP Component as Semantic Network Error: Insights from a Feature-Based Connectionist Attractor Model of word Meaning". In: *Cognition* 132.1 (2014), pp. 68–89 (cit. on p. 19).

[RB12] Stéphane Rauzy and Philippe Blache. "Robustness and Processing Difficulty Models. A Pilot Study for Eye-Tracking Data on the French Treebank". In: *Proceedings of the COLING Workshop on Eye-Tracking and NLP*. 2012 (cit. on pp. 11, 15).

[Rim+16] Laura Rimell, Jean Maillard, Tamara Polajnar, and Stephen Clark. "RELPRON: A relative clause evaluation data set for compositional distributional semantics". In: *Computational Linguistics* 42.4 (2016), pp. 661–701 (cit. on pp. 70, 99, 101).

[RG65] Herbert Rubenstein and John B Goodenough. "Contextual Correlates of Synonymy". In: *Communications of the ACM* 8.10 (1965), pp. 627–633 (cit. on p. 61).

[RG10] Sebastian Rudolph and Eugenie Giesbrecht. "Compositional Matrix-Space Models of Language". In: *Proceedings of the ACL*. 2010 (cit. on p. 66).

[Rum79] David E Rumelhart. “Some Problems with the Notion of Literal Meanings”. In: *Metaphor and Thought* 72 (1979) (cit. on p. 33).

[Sag12] Ivan A Sag. “Sign-Based Construction Grammar: An Informal Synopsis”. In: *Sign-Based Construction Grammar* 193 (2012), pp. 69–202 (cit. on pp. 21, 77).

[Sah08] Magnus Sahlgren. “The Distributional Hypothesis”. In: *Italian Journal of Linguistics* 20 (2008), pp. 33–53 (cit. on pp. 46, 47).

[SHK08] Magnus Sahlgren, Anders Holst, and Pentti Kanerva. “Permutations as a Means to Encode Order in Word Space”. In: *Proceedings of the Annual Conference of the Cognitive Science Society*. 2008 (cit. on p. 39).

[SL16] Magnus Sahlgren and Alessandro Lenci. “The Effects of Data Size and Frequency Range on Distributional Semantic Models”. In: *Proceedings of EMNLP*. 2016 (cit. on pp. 53, 54, 57).

[San+17] Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. “Measuring Thematic Fit with Distributional Feature Overlap”. In: *Proceedings of EMNLP*. 2017 (cit. on pp. 50, 78, 89, 125).

[San+16a] Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, Chu-Ren Huang, and Philippe Blache. “Testing APsyn against Vector Cosine on Similarity Estimation”. In: *Proceedings of PACLIC*. 2016 (cit. on p. 53).

[San+16b] Enrico Santus, Tin-Shing Chiu, Qin Lu, Alessandro Lenci, and Chu-Ren Huang. "What a Nerd! Beating Students and Vector Cosine in the ESL and TOEFL Datasets". In: *Proceedings of LREC*. 2016 (cit. on p. 50).

[San+18] Enrico Santus, Hongmin Wang, Emmanuele Chersoni, and Yue Zhang. "A Rank-Based Similarity Metric for Word Embeddings". In: *Proceedings of ACL*. 2018 (cit. on p. 50).

[SD14] Asad Sayeed and Vera Demberg. "Combining Unsupervised Syntactic and Semantic Models of Thematic Fit". In: *Proceedings of CLIC.it*. 2014 (cit. on pp. 58, 64, 125).

[SDS15] Asad Sayeed, Vera Demberg, and Pavel Shkadzko. "An Exploration of Semantic Features in an Unsupervised Thematic Fit evaluation Framework". In: *Italian Journal of Computational Linguistics* 1.1 (2015) (cit. on pp. 58, 64, 78, 125).

[SFD15] Asad Sayeed, Stefan Fischer, and Vera Demberg. "Vector-Space Calculation of Semantic Surprisal for Predicting Word Pronunciation Duration". In: *Proceedings of ACL*. 2015 (cit. on p. 15).

[SGD16] Asad Sayeed, Clayton Greenberg, and Vera Demberg. "Thematic Fit Evaluation: An Aspect of Selectional Preferences". In: *Proceedings of the ACL Workshop on Evaluating Vector Space Representations for NLP*. 2016 (cit. on pp. 58, 63).

[SA13] Roger C Schank and Robert P Abelson. *Scripts, Plans, Goals, and Understanding: An inquiry into Human Knowledge Structures*. Psychology Press, 2013 (cit. on pp. 37, 76).

[Shu09] Ekaterina Shutova. "Sense-Based Interpretation of Logical Metonymy Using a Statistical Method". In: *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*. Suntec, Singapore, 2009, pp. 1–9 (cit. on p. 107).

[Smo90] Paul Smolensky. "Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems". In: *Artificial intelligence* 46.1-2 (1990), pp. 159–216 (cit. on pp. 62, 66).

[Soc+11] Richard Socher, Eric H Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y Ng. "Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection". In: *Advances in Neural Information Processing Systems*. 2011 (cit. on p. 68).

[Soc+12] Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. "Semantic Compositionality through Recursive Matrix-Vector Spaces". In: *Proceedings of EMNLP-CONLL*. 2012 (cit. on p. 70).

[Søg16] Anders Søgaard. "Evaluating Word Embeddings with fMRI and Eye-Tracking". In: *Proceedings of the ACL Workshop on Evaluating Vector-Space Representations for NLP* (2016) (cit. on p. 62).

[SW86] Dan Sperber and Deirdre Wilson. *Relevance: Communication and Cognition*. Oxford: Blackwell, 1986 (cit. on pp. 82, 106).

[Spi96] Michael J Spivey-Knowlton. "Integration of Visual and Linguistic Information: Human Data and Model Simulations". PhD thesis. University of Rochester, 1996 (cit. on p. 25).

[Til+16] Ottokar Tilk, Vera Demberg, Asad Sayeed, Dietrich Klakow, and Stefan Thater. "Event Participant Modelling with Neural Networks". In: *Proceedings of the EMNLP*. 2016 (cit. on pp. 65, 125, 126, 129, 130).

[TPM02] Matthew J Traxler, Martin J Pickering, and Brian McElree. "Coercion in Sentence Processing: Evidence from Eye-Movements and Self-Paced Reading". In: *Journal of Memory and Language* 47.4 (2002), pp. 530–547 (cit. on pp. 104, 105, 107–109, 113, 119).

[Tur12] Peter D Turney. "Domain and Function: A Dual-Space Model of Semantic Relations and Compositions". In: *Journal of Artificial Intelligence Research* 44 (2012), pp. 533–585 (cit. on pp. 59, 129).

[TP10] Peter D Turney and Patrick Pantel. "From Frequency to Meaning: Vector Space Models of Semantics". In: *Journal of Artificial Intelligence Research* 37 (2010), pp. 141–188 (cit. on pp. 11, 49, 51, 52).

[Van+05a] Jos JA Van Berkum, Colin M Brown, Pienie Zwitserlood, Valesca Kooijman, and Peter Hagoort. "Anticipating Upcoming Words in Discourse: Evidence from ERPs and Reading Times." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31.3 (2005), p. 443 (cit. on p. 19).

[Van10] Tim Van de Cruys. "A Non-Negative Tensor Factorization Model for Selectional Preference Induction". In: *Natural Language Engineering* 16.4 (2010), pp. 417–437 (cit. on p. 66).

[VK83] Teun A Van Dijk and Walter Kintsch. *Strategies of Discourse Comprehension*. Academic Press, 1983 (cit. on p. 38).

[Van+05b] Roger PG Van Gompel, Martin J Pickering, Jamie Pearson, and Simon P Liversedge. "Evidence Against Competition During Syntactic Ambiguity Resolution". In: *Journal of Memory and Language* 52.2 (2005), pp. 284–307 (cit. on p. 23).

[Vas03a] Shravan Vasishth. "Quantifying Processing Complexity in Human Sentence Parsing: The Role of Decay, Activation and Similarity-Based Interference". In: *Proceedings of EuroCogSci*. 2003 (cit. on p. 13).

[Vas03b] Shravan Vasishth. *Working Memory in Sentence Comprehension: Processing Hindi Center Embeddings*. Routledge, 2003 (cit. on p. 13).

[Vas+18] Paolo Vassallo, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. "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)*. Miyazaki, Japan, 2018 (cit. on p. 126).

[Vec+17] Eva M Vecchi, Marco Marelli, Roberto Zamparelli, and Marco Baroni. "Spicy Adjectives and Nominal Donkeys: Capturing Semantic Deviance Using Compositionality in Distributional Spaces". In: *Cognitive science* 41.1 (2017), pp. 102–136 (cit. on p. 124).

[WZC17] Zhongqing Wang, Yue Zhang, and Ching-Yun Chang. "Integrating Order Information and Event Relation for Script Event Prediction". In: *Proceedings of EMNLP*. 2017 (cit. on p. 130).

[War+15] Tessa Warren, Evelyn Milburn, Nikole D Patson, and Michael Walsh Dickey. "Comprehending the Impossible: What Role Do Selectional Restriction Violations Play?" In: *Language, cognition and neuroscience* 30.8 (2015), pp. 932–939 (cit. on pp. 22, 42, 127).

[WBC18] Noah Weber, Niranjan Balasubramanian, and Nathanael Chambers. "Event Representations with Tensor-based Compositions". In: *Proceedings of AAAI*. 2018 (cit. on pp. 129, 130).

[WK99] Jill Weckerly and Marta Kutas. "An Electrophysiological Analysis of Animacy Effects in the Processing of Object Relative Sentences". In: *Psychophysiology* 36.5 (1999), pp. 559–570 (cit. on p. 42).

[Wid08] Dominic Widdows. "Semantic Vector Products: Some Initial Investigations". In: *Second AAAI Symposium on Quantum Interaction*. Vol. 26. 2008 (cit. on p. 66).

[Zan+10] Fabio Massimo Zanzotto, Ioannis Korkontzelos, Francesca Fallucchi, and Suresh Manandhar. "Estimating Linear Models for Compositional Distributional Semantics". In: *Proceedings of COLING*. 2010 (cit. on p. 68).

[Zar+13] Alessandra Zarcone, Alessandro Lenci, Sebastian Padó, and Jason Utt. "Fitting, not Clashing! A Distributional Semantic Model of Logical Metonymy". In: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*.

Potsdam, Germany, 2013, pp. 404–410 (cit. on pp. 107, 111, 117, 119).

[ZP11] Alessandra Zarcone and Sebastian Padó. "Generalized Event Knowledge in Logical Metonymy Resolution". In: *Proceedings of the 33rd annual meeting of the Cognitive Science Society (CogSci 2011)*. Boston, MA, 2011 (cit. on p. 106).

[ZPL14] Alessandra Zarcone, Sebastian Padó, and Alessandro Lenci. "Logical Metonymy Resolution in a Words-as-Cues Framework: Evidence from Self-paced Reading and Probe Recognition". In: *Cognitive Science* 38.5 (2014), pp. 973–996 (cit. on pp. 105, 106).

[ZUP12] Alessandra Zarcone, Jason Utt, and Sebastian Padó. "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)*. Montréal, Canada, 2012, pp. 70–79 (cit. on pp. 107, 110, 112, 119).

[ZM05] Rolf A Zwaan and Carol J Madden. "Embodied Sentence Comprehension". In: *Grounding cognition: The role of perception and action in memory, language, and thinking* (2005), pp. 224–245 (cit. on pp. 38, 81).

[ZR98] Rolf A Zwaan and Gabriel A Radvansky. "Situation Models in Language Comprehension and Memory." In: *Psychological Bulletin* 123.2 (1998), p. 162 (cit. on pp. 38, 81).

ANNEXES

A. Intitulés des doctorats AMU

Mentions et Spécialités des doctorats votées en CS le 16/10/2012

ED 62 SCIENCES DE LA VIE ET DE LA SANTE

- Biologie
  - Biochimie structurale
  - Génomique et Bioinformatique
  - Biologie du développement
  - Immunologie
  - Génétique
  - Microbiologie
  - Biologie végétale
- Neurosciences
- Pathologie humaine
  - Oncologie
  - Maladies infectieuses
  - Génétique humaine
  - Conseil en Génétique
  - Pathologie vasculaire et nutrition
  - Ethique
  - Recherche clinique et Santé Publique

ED 67 SCIENCES JURIDIQUES ET POLITIQUES

— Droit privé

— Droit public

— Histoire du droit

— Droit

— Science politique

ED 184 MATHEMATIQUES ET INFORMATIQUE

— Mathématiques

— Informatique

— Automatique

ED 250 SCIENCES CHIMIQUES DE MARSEILLE

— Sciences chimiques

ED 251 SCIENCES DE L'ENVIRONNEMENT

— Anthropologie biologique

— Ecologie

— Géosciences de l'environnement

— Génie des procédés

— Océanographie

— Chimie de l'environnement

ED 352 PHYSIQUE ET SCIENCES DE LA MATIERE

— Astrophysique et Cosmologie

— Biophysique

— Energie, Rayonnement et Plasma

— Instrumentation

— Optique, Photonique et Traitement d'Image

— Physique des Particules et Astroparticules

— Physique Théorique et Mathématique

— Matière Condensée et Nanosciences

ED 353 SCIENCES POUR L'INGENIEUR: MECANIQUE, PHYSIQUE, MICRO ET NANOELECTRONIQUE

— Energétique

— Mécanique et Physique des Fluides

— Acoustique

— Mécanique des Solides

— Micro et Nanoélectronique

— Génie Civil et Architecture

ED 354 LANGUES, LETTRES ET ARTS

— Etudes anglophones

— Etudes germaniques

— Etudes slaves

— Langue et littérature chinoises

— Langue et Littérature françaises

— Littérature générale et comparée

— Arts plastiques et sciences de l'Art

— Musicologie

— Etudes cinématographiques

— Arts du spectacle

ED 355 ESPACES, CULTURES, SOCIETES

— Géographie

— Urbanisme et Aménagement du territoire

— Préhistoire

— Archéologie

— Histoire de l'Art

— Histoire

— Sciences de l'Antiquité

— Mondes arabe, musulman et sémitique

— Etudes romanes

— Sociologie

— Anthropologie

— Architecture

ED 356 COGNITION, LANGAGE, EDUCATION

- Philosophie
- Psychologie
- Sciences du Langage
- Sciences de l'Information et de la Communication
- Sciences de l'Education

ED 372 SCIENCES ECONOMIQUES ET DE GESTION

- Sciences de Gestion
- Sciences Economiques
- Sciences Economiques: AMSE

ED 463 SCIENCES DU MOUVEMENT HUMAIN

- Sciences du Mouvement Humain
  - Biomécanique
  - Contrôle Perceptivo-Moteur et Apprentissage
  - Physiologie de l'exercice
  - Sciences de l'Homme et de la Société