6 A Distributional Model of Verb-Specific Semantic Roles Inferences

Gianluca E. Lebani and Alessandro Lenci

Abstract

In a standard view, commonly adopted in psycholinguistics and computational linguistics, thematic roles are approached as primitive entities able to represent the roles played by the arguments of a predicate. In theoretical linguistics, however, the inability to reach a consensus on a primitive set of semantic roles led to the proposal of new approaches in which thematic roles are better described as a bundle of more primitive entities (e.g., Dowty, 1991; Van Valin, 1999) or as structural configurations (e.g., Jackendoff, 1987). In a complementary way, psycholinguistic evidence supports the idea that thematic roles and nominal concepts are represented in similar ways (McRae et al., 1997b; Ferretti et al., 2001), thus suggesting that the former can be accounted for as predicate-specific bundles of inferences activated by the semantics of the verb (e.g., the patient of *kill* is typically alive before the event and dead afterward). Such inferences can take the form of either presuppositions or entailment relations activated when a filler saturates a specific argument position for a given predicate.

Our aim in this chapter is twofold. First, we report behavioral data collected to obtain a more fine-grained characterization of the thematic role properties activated by a subset of English verbs. To this end, we employed the modified version of the McRae et al. (1997b) elicitation paradigm proposed by Lebani et al. (2015) to describe which semantic properties of the participants are more relevant in each phase of the action described by the predicate. Next, we test the possibility to model such verb-specific inference patterns by exploiting corpus-based distributional data, thus proposing a novel approach to represent the same level of semantic knowledge that is currently described by means of a finite set of thematic roles.

6.1 Representing and Acquiring Thematic Roles

The concept of *thematic role* is one of the most vaguely defined, yet appealing, technical tools in the linguist's toolkit. Since its settlement in the circle of relevant modern theoretical issues, thanks to investigations by Tesnière (1959), Gruber (1965), Fillmore (1968), and Jackendoff (1972), this concept has been approached with what Dowty (1989) called the *I-can't-define-it-but-I-know-it-when-I-see-it* stance; that is, by using it without offering a proper definition. As easily predictable, such a state of affairs led to the proliferation of many alternative terms forged to refer to very close, if not identical, intuitions: *case relations*, *theta roles*, *semantic roles* and *thematic relations*. All these approaches share the general idea that thematic roles describe what can be intuitively depicted as the role played by an argument in the event or situation described by a verb, and little formalization has been obtained since early documented proposals such as Pānini's *kārakas*.

In natural language processing (NLP), thematic roles are both a valuable source of semantic knowledge encoded in lexical resources such as VerbNet (Kipper-Schuler, 2005; Kipper et al., 2008), FrameNet (Baker et al., 1998), and PropBank (Kingsbury and Palmer, 2003), as well as the target of automatic extraction models, usually referred to as Semantic Role Labeling tools (see Gildea and Jurafsky, 2002; Lluís Màrquez, 2008; Palmer et al., 2010). This information has proven useful for a variety of tasks, including machine translation (e.g., Liu and Gildea, 2010; Wu and Palmer, 2011) and question answering (e.g., Shen and Lapata, 2007). However, most of the computational linguistics literature still sees semantic roles as unanalyzable and unitary entities, thus relying on a view whose dramatic limitations have long been identified (for a review, see Levin and Rappaport Hovav, 2005).

A real breakthrough in the linguistic research on thematic roles was carried out by David Dowty. Rather than pursuing the impossible goal of defining an exhaustive taxonomy of semantic roles, Dowty argued that roles are not discrete and categorical entities, but have the same prototype structure of other types of concepts. Dowty (1989) proposed a *neo-Davidsonian* approach in which thematic roles are seen as a set of entailments of a predicate over its arguments, and thus characterized as second-order properties, i.e., as predicates of predicates. He also distinguished *individual roles* from *linguistic roles*. The former are verb-specific roles defined by the entailments associated with a particular verb argument: for instance, the *builder-role* is the set of all the properties and inferences we can conclude about *x* solely from knowing that *x builds y* is true. Linguistic roles are instead more abstract concepts shared among many verbs. Dowty (1991) assumed two basic linguistic roles, proto-agent and proto-patient, defined as a clusters of properties or entailments, organized like

the prototypes of Rosch and Mervis (1975). For instance, he described the proto-agent role as characterized by entailments such as ⟨*volitional involvement in the event*⟩, ⟨*sentience and/or perception*⟩, etc. Linguistic roles take on a special status in linguistic theory as they enter into grammatical generalizations, given that proto-agents and proto-patients tend to be realized as subjects and direct objects, respectively, in active sentences.

On the NLP side, work by Reisinger et al. (2015) shows that Dowty's proto-role hypothesis can be empirically validated by exploiting a large-scale crowd-sourced annotation task using corpus data. These authors then compared the results of the annotation they collected against those available in more conventional resources such as VerbNet. By building on encouraging results, finally, these scholars propose a novel task, Semantic Proto-Role Labeling, in which a system is asked to annotate a sentence with "scalar judgments of Dowty-inspired properties," rather than with more conventional categorical thematic roles.

Acknowledging that thematic roles, both verb-specific and general, are to be conceived as clusters of properties entailed by verb arguments in turn raises two crucial issues that represent the main focus of this paper: (1) *How can we identify the specific entailments that characterize the thematic roles of a verb?* (2) *How do we learn the entailments associated with these thematic roles?* The first issue concerns the empirical evidence we can use to ground the study of thematic roles on a firm scientific foundation. McRae et al. (1997b) propose to identify the entailments of verb-specific roles using the features produced by a group of native speakers in a norming experiment. The feature-norming paradigm is in fact commonly adopted to investigate the content of conceptual knowledge in semantic memory. Because thematic role concepts are conceived as clusters of properties, subjects' elicited features can be used to identify information associated with the roles of specific events and to estimate its degree of prototypicality.

Concerning the way thematic roles are acquired, we endorse the claim by McRae et al. (1997b) that "role concepts are formed through the everyday experiences during which people learn about the entities and objects that tend to play certain roles in certain events" (p. 141). Similar to nominal concepts, thematic roles are organized in hierarchical structures leading from verb-specific roles to more abstract thematic concepts: for instance, the role of "accuser" is regarded as a subtype of the more general role of "agent." Therefore, both individual and general roles result from an inductive process of abstraction from event knowledge. This is similar to the way roles are organized in FrameNet: verbs evoke event-specific frames that are part of an inheritance network whose top nodes correspond to abstract event schemas containing general roles like agent or patient. Psycholinguistic research has indeed provided robust evidence that

online sentence processing is deeply influenced by knowledge about events and their thematic roles (for a review, see McRae and Matsuki, 2009): verbs seem to be able to prime nouns describing the typical participant to the event they describe (Altmann and Kamide, 1999; Ferretti et al., 2001; Hare et al., 2009), especially in the presence of certain syntactic and grammatical cues (Traxler et al., 2001; Ferretti et al., 2001, 2007; Altmann and Kamide, 2007); nouns too appear to be able to prime both the other participants of an event (McRae et al., 1998; Kamide et al., 2003; Bicknell et al., 2010), as well as those verbs describing the events in which they typically participate (McRae et al., 2005a), a behavior that is useful to select a given verb sense (Matsuki et al., 2011). This knowledge is referred to by McRae and Matsuki (2009) as *Generalized Event Knowledge* because it consists of general encyclopedic information about the prototypical organization and unfolding of events. Generalized Event Knowledge is acquired through different sources, most importantly first-hand participation in events and language. For instance, the entailments characterizing the agent role of *accuse* derive from our experiences with people who accuse others and from linguistic descriptions of such events.

Our goal in this chapter is to investigate the contribution of language as a source of the entailments that characterize verb-specific semantic roles. In particular, we aim to explore to what extent the entailments activated by the thematic roles of a subset of English verbs can be acquired from their usage in a corpus. To address this question, we are proposing a distributional model in which the semantic content of the proto-agent and proto-patient role of a verb are characterized by the sets of verbs and nominal predicates that are strongly associated with them in texts. As an example, what our model looks after is the fact that the agent role of the target verb TO EAT can be described with properties such as *s/he drinks (while eating)*, that *s/he will digest what s/he has eaten*, and that *s/he was previously hungry*. In this preliminary work, this information will be represented by a strong association of this verb-specific role with verbs like ⟨*to drink*⟩ and ⟨*to digest*⟩, as well as with adjectives like ⟨*hungry*⟩.[1] We will test our model by comparing the extracted information against the properties produced by a group of native speakers to describe the content of verb-specific thematic roles.

The chapter is organized as follows. In Section 6.2 we illustrate a feature-norming study by means of which, following works by McRae et al. (1997b) and Lebani et al. (2015), we describe the thematic roles associated with twenty English verbs. In Sections 6.3 and 6.4 we show how a simple distributional model is apt to extract such information from a corpus, albeit with

[1] Throughout these pages, target verbs will be printed in SMALL CAPITAL font, whereas speaker-generated and automatically extracted descriptions will be enclosed in ⟨*angles brackets*⟩.

a series of limitations and blindspots on which we will speculate in the final section.

6.2 Characterizing the Semantic Content of Verb Proto-roles

In the modern psycholinguistic literature, the feature norm paradigm has been widely employed to characterize the semantic content of the human conceptual knowledge. In its simpler form, it requires native speakers to produce short phrases to describe a set of target concepts. The collected descriptions are then normalized and categorized by the experimenter to build a dataset of pairings concept-feature of the form DOG ⟨*has a tail*⟩, LOUNGE ⟨*is fancy*⟩, or AIRPLANE ⟨*flies*⟩.

Freely available resources built by exploiting different implementations of the feature norm paradigm are available for a limited number of languages, including English (Garrard et al., 2001; McRae et al., 2005b; Vinson and Vigliocco, 2008; Devereux et al., 2014), Italian (Kremer and Baroni, 2011; Lebani, 2012; Montefinese et al., 2013; Lenci et al., 2013), Dutch (De Deyne et al., 2008), and German (Kremer and Baroni, 2011; Roller and Schulte im Walde, 2014). These collections have been used as experimental stimuli (Ashcraft, 1978; Vigliocco et al., 2006), as a source of knowledge in proposing a model of semantic memory (Collins and Loftus, 1975; Hinton and Shallice, 1991; McRae et al., 1997a; Vigliocco et al., 2004; Storms et al., 2010), to investigate the pattern of impairments shown by anomic patients (Garrard et al., 2001; McRae and Cree, 2001; Vinson et al., 2003; Sartori and Lombardi, 2004), and in research on the nature of empirical phenomena such as semantic priming (Cree et al., 1999; Vigliocco et al., 2004), semantic compositionality (Hampton, 1979), and categorization (Smith et al., 1974; Rosch and Mervis, 1975).

In the computational linguistics literature, feature norms collections have been used to evaluate semantic extraction methods (Baroni et al., 2008; Baroni and Lenci, 2010) or as a source of semantic knowledge that can be exploited to enrich existing resources or other kinds of knowledge (Barbu and Poesio, 2008; Andrews et al., 2009; Steyvers et al., 2011; Lebani, 2012; Fagarasan et al., 2015). Some scholars even tested systems specifically tuned to extract feature-like semantic knowledge (Poesio et al., 2008; Devereux et al., 2009; Baroni et al., 2010; Kelly et al., 2010, 2013).

With few exceptions, most of the available feature norms have been collected for nominal concrete concepts expressed as nominal entities. One of these exceptions is the dataset assembled by Vinson and Vigliocco (2008), where 287 of the 456 described concepts denote actions, in 217 cases by means of a verbal lemma. In the paradigm adopted by these scholars there is no difference in the way verbs and nouns are collected and represented, so that the final dataset

represents a unitary space, whose suitability for modeling the human semantic memory has been proved by the same authors (Vigliocco et al., 2004).

Whereas Vinson and Vigliocco (2008) were interested in the properties of the event denoted by the verb, McRae et al. (1997b) collected the characteristics of the proto-agent and proto-patient roles for a group of 20 English transitive verbs, thus showing how the feature norm paradigm can be used to empirically characterize the semantic content of thematic roles. The scholars opted for a traditional paper-and-pencil setting, in which 32 participants were asked to list the characteristics of only one role for each verb. Crucially, instructions explicitly stated that what they were asked to list were not the typical fillers of a role (e.g., *judge* as the agent of TO CONVICT), but their characteristics (e.g., ⟨*is old*⟩ for the same role). McRae et al. (1997b) collected 1,573 distinct descriptions, 445 of which have been produced by 3 or more participants. Overall, no clear advantage of one proto-role over the other has been recorded, but the distribution of the features in the different verb-role pairs is far from uniform, a phenomenon that the authors ascribed to the well-known fact that some roles for some verbs admit a more restrictive group of fillers than others. Examples of highly consistent verb-specific roles include the proto-agent of the verb TO RESCUE and the proto-patient of the verb TO TEACH, whereas loosely defined roles include the patients of TO ACCUSE and TO SERVE.

By building on the observations by McRae and colleagues, Lebani et al. (2015) applied a modified version of this paradigm to a set of 20 Italian verbs. These authors modified the original methodology in several ways: by submitting the questionnaire online to a group of selected participants; by asking each participant to rate all possible verb-role pairs; by providing the participants with instructions to the form "*describe who* CONVICTS" or "*describe who* IS CONVICTED", to avoid confronting the subjects with an elusive concept such as thematic role. The biggest modification to the original paradigm, however, was the explicit request to describe each role of each verb with respect to three different time slots:

- *before* the event described by the verb takes place: for instance, properties like ⟨*to be ill*⟩ for the patient of the verb TO CURE;
- *while* the event described by the verb takes place: for instance, properties like ⟨*to speak*⟩ for the agent of the verb TO TEACH;
- *after* the event described by the verbs took place: for instance, properties like ⟨*to feel fine*⟩ for the patient of the verb TO CURE.

Lebani et al. (2015) evaluated the impact of this last manipulation against an online reimplementation of McRae's paradigm (McRae et al., 1997b), and the collected features set was much less skewed toward the required characteristics, and way more informative of the entailed properties that a filler acquires

by participating in the event described by a verb. It is this characteristic that drove our choice to adopt this last paradigm to collect, for 20 English transitive verbs, a description of the semantic content of the proto-agent and proto-patient semantic roles, to be later used as an evaluation benchmark against the neo-Davidsonian distributional model that we describe in the next section.

6.2.1 Method

To collect data from English native speakers, we crowdsourced our elicitation task through the Crowdflower marketplace.[2] Such a solution is usually adopted to collect a great amount of annotations or data, and to do so as quickly and cheaply as possible. But this often comes with the price of lower reliability and/or precision of the data due to the influence of many noncontrollable variables (on these topics, see Snow et al., 2008; Fort et al., 2011). Even if other authors proved that the collection of featural descriptions is a task that can be easily crowdsourced (e.g., Roller and Schulte im Walde, 2014), this required an adaptation of the procedure in Lebani et al. (2015) in order to submit our workers to a task that is not too labor-intensive and to filter unreliable data (see Kittur et al., 2013).

Materials We borrowed our experimental stimuli from McRae et al. (1997b). These were the following 20 English transitive verbs holding animate agents and patients: TO CONVICT, TO TEACH, TO RESCUE, TO ENTERTAIN, TO FIRE, TO CURE, TO PUNISH, TO HIRE, TO EVALUATE, TO ARREST, TO LECTURE, TO FRIGHTEN, TO INSTRUCT, TO TERRORISE, TO INVESTIGATE, TO WORSHIP, TO INTERVIEW, TO ACCUSE, TO SERVE, TO INTERROGATE.

The semantics of each possible verb-role-slot combination was then paraphrased to create requests of the form "please, list some of the features possessed by someone that [*inflected verb*] someone else" for the agent role and "please, list some of the features possessed by someone that is [*inflected verb*] by someone else." For instance, the six requests created for the agent and patient role of the verb TO FIRE were, respectively: "please, list some of the features possessed by someone that [*fired* | *is firing* | *is going to fire*] someone else" and "please, list some of the features possessed by someone that [*has been fired* | *is being fired* | *is going to be fired*] by someone else". Overall, 120 test questions of this sort were created, each to be used as the microtask to be submitted to our workers.

Procedure In each microtask the worker was requested to supply 5 to 10 short descriptions for a verb-role-slot triple. Microtasks were submitted

[2] Accessible at www.crowdflower.com

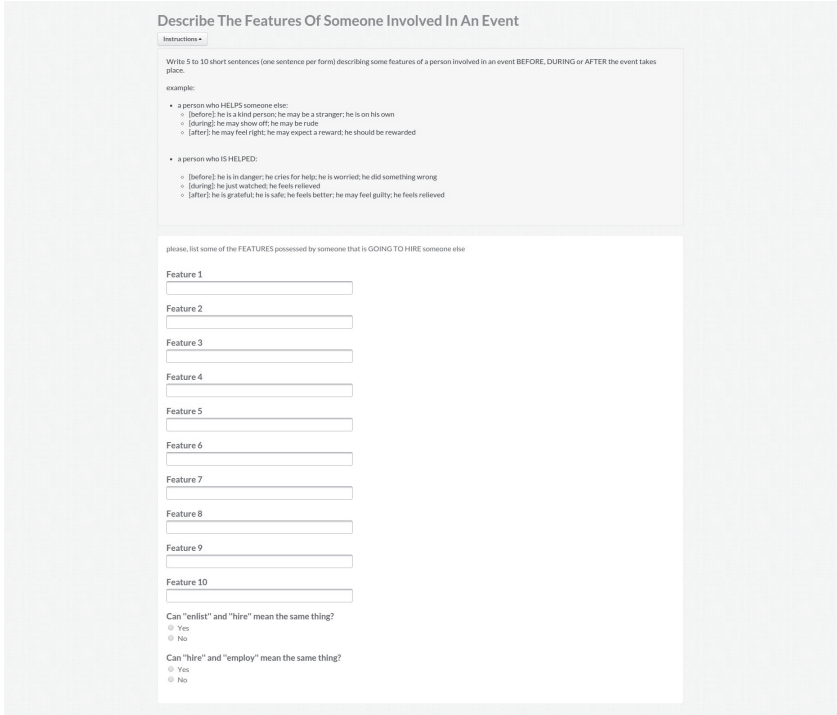

Figure 6.1 The verb role description interface in Crowdflower.

by means of a web page similar to the one in Figure 6.1. The top of the page supplied intuitive instructions, along with exemplar descriptions for the verb TO HELP. The main area of the page, i.e., the one with a white background, presented the test question, followed by 10 empty forms and 2 language comprehension questions. Test questions required the worker to indicate whether the meaning of the target verb was similar to that of a test verb, which could be either a synonym of the target or a completely unrelated word. Each worker was free to complete from 1 to 120 different microtasks, presented in randomized order. On average, workers needed 116.02 s (SD = 96.98) to complete a valid hit. Hits were rejected if they met any of the following conditions:

- the worker didn't answer correctly to any test question;
- the worker completed the task in less than 30 s[3];

[3] Both the use of test questions and the duration threshold were intended to identify scammers. As a matter of fact, a manual inspection of the data showed that the latter strategy was more efficient than the former.

- the worker failed to provide at least three valid descriptions;
- the worker clearly misunderstood the requirements of the task.

The data collection process took place at the end of September 2014, and ended when 15 usable annotations for each verb-role-slot question was recorded, that is, after approximately 7 days.

Participants Eighty-seven unique workers contributed to the norming experiment, receiving €0.05 per hit. Only Crowdflower-certified "highest quality" contributors from the United Kingdom, the United States, or Ireland were allowed to participate. On average, each subject completed 20.7 (SD = 26.64) approved hits.

6.2.2 Selection and Normalization

The collected raw descriptions were first manually inspected to remove unwanted material such as incomplete sentences, meaningless descriptions, and all cases in which the worker reported the filler of the thematic role rather than its characteristics.

The selected description were then "normalized," that is, manipulated to identify meaningful chunks of information. Normalization practices in the literature can be organized into three main classes: minimal normalization (e.g., De Deyne et al., 2008); raw descriptions rewritten to conform to a phrase template (e.g., McRae et al., 1997b; Garrard et al., 2001; McRae et al., 2005b; Kremer and Baroni, 2011; Lenci et al., 2013; Devereux et al., 2014; Roller and Schulte im Walde, 2014; Lebani et al., 2015); or raw descriptions reduced to a list of focal concepts (e.g., Vinson and Vigliocco, 2008; Lebani, 2012). Common to virtually all strategies is a first step in which:

- spelling and orthography are standardized;
- conjoint and disjunct features are split: accordingly, a description such as ⟨*is tasty and delicious*⟩ should be split into ⟨*is tasty*⟩ and ⟨*is delicious*⟩;
- auxiliaries and modal are stripped away: for instance, the description ⟨*could be guilty*⟩ should be simplified into a phrase like ⟨*is guilty*⟩.

The main reason for us to collect featural descriptions was to evaluate a distributional model, so that the subsequent normalization steps were aimed at a – sometimes brutal – reduction of the raw description phrases into lists of focal concepts, a strategy analogous to those adopted by Vinson and Vigliocco

(2008) and Lebani (2012).[4] In this second step, several crucial manipulations are performed:

- quantifiers are removed: for instance, the description ⟨*has five legs*⟩ can be simplified into something of the form ⟨*has legs*⟩;
- the prominent concept(s) of each description are identified, and the remaining linguistic material discarded: for instance, two important chunks of information are available in the description ⟨*has beautiful legs*⟩, thus leading to the creation of the two focal features ⟨*beautiful*⟩ and ⟨*legs*⟩;
- the identified focal concepts are then lemmatized: e.g., plural nouns become singular, participles and gerunds are reported in their base form.
- synonymous features produced in different hits were encoded by using their most recurrent linguistic form: if two workers produced the description ⟨*is calm*⟩ and another produced ⟨*is cool*⟩, then all descriptions were treated as instances of the same feature, i.e., ⟨*is calm*⟩. Synonymous descriptions or focal concepts produced in the same hit, however, were analyzed as redundancies and discarded.

Slots merging and norms expansion The dataset of verb-role-slot features collected so far is analogous to the one described by Lebani et al. (2015), and throughout this chapter we refer to its features as *slot-based features*. For two reasons, however, this dataset is not optimal for our purposes, that is, to serve as a gold standard in the evaluation of our model. First of all, our model does not attempt to extract the temporal signature of each feature. The reason we resorted to this paradigm was its superiority in extracting "entailed" properties. We therefore merged all the feature sets produced for a given verb-role pair, irrespective of their temporal characterization. We refer to these as *role-based features*.

The second issue has been recognized and widely discussed in the relevant literature (e.g., Barbu and Poesio, 2008; Baroni et al., 2008; Baroni and Lenci, 2010). It pertains to the fact that the normalization process has the side effect of reducing the lexical richness of the uttered descriptions. When using a feature norm collection as a gold standard, lexical paucity has a direct impact on the evaluation statistics by artificially increasing the number of false negatives (i.e., properties extracted by the system but not linked to a synonymic description in the norms). In using the concrete concept properties of McRae and colleagues (McRae et al., 2005b) as a gold standard for the European

[4] In fairness, different sets of norms have been prepared, each developed by following one of the three different normalization strategies. Given the scope of this chapter in these pages, we focus solely on those obtained by reducing the raw descriptions to their focal concepts.

Summer School in Logic, Language, and Information (ESSLLI) 2008 Distributional Semantic Workshop unconstrained property-generation task, Baroni et al. (2008) expanded their reference norms by (1) selecting the top ten features for each described concept, (2) extracting from WordNet (Fellbaum, 1998) the synonyms of each last word of each feature, and (3) performing a manual check to filter irrelevant synonyms and to add other potential linguistic material. Along these lines, we expanded our role-based features by extracting from WordNet (Fellbaum, 1998) all the synonyms of each of our focal concepts, without manual intervention. We refer to these as *expanded-role-based features*.

6.2.3 Results

Overall, our workers produced 11,985 raw descriptions, uniformly distributed along thematic roles (6,066 for the agent roles and 5,918 for the patient roles) and time slots (3,964 for the *before* slots, 4,016 for the *during* slots, and 4,004 for the *after* slots). Each hit returned, on average, 6.66 raw features (SD = 2.17). By splitting conjoint and disjunct descriptions the total climbs to 12,091, of which 392 were later discarded because they contained unwanted material or redundant information.

The normalization process resulted in 12,802 raw slot-based features. From these, 9,667 distinct verb-role-slot features were collected: 5,136 for the agent roles and 4,531 for the patient ones. In contrast with that reported by Lebani et al. (2015), this difference reaches statistical significance according to a paired Student's *t*-test ($t = 7.49$, $df = 19$, $p < 0.001$). On the other side, the distribution is pretty even across the different time slots: 3,190, 3,272, and 3,205 for the *before*, *during*, and *after* slots, respectively. A chi-square analysis failed to reveal any significant pattern both in the distribution of the features for slot both in the whole dataset ($\chi^2 = 0.57$, $df = 2$, $p > 0.1$), and among the two groups of thematic roles ($\chi^2 = 1.18$, $df = 2$, $p > 0.1$).

On average, each distinct slot-based feature has been produced by 1.32 (SD $= 0.945$) workers, and consistent features (those with frequency ≥ 2) accounts for the 17.69% of the total distinct slot-based features: 827 for the agent roles and 883 for the patient ones; 572, 563, and 575 for the *before*, *during*, and *after* slots, respectively. A paired Student's *t*-test shows that the difference in consistency between the two thematic roles reaches statistical significance, too ($t = -5.88$, $df = 19$, $p < 0.001$). A chi-square analysis failed to reveal any significant pattern both in the feature consistency of the time slots both in the whole dataset ($\chi^2 = 0.34$, $df = 2$, $p > 0.1$) and among the two groups of thematic roles ($\chi^2 = 0.14$, $df = 2$, $p > 0.1$).

Table 6.1 reports the number of featural descriptions and their consistency across the verb-role pairings. What it clearly shows is that, abstracting away

Table 6.1 *Distinct Slot-Based Features and Consistency for Verb-Role Pair*

	Agent		Patient	
	SB features	Consistency[a]	SB features	Consistency
TO ACCUSE	275	9.09%	214	19.63%
TO ARREST	259	18.15%	237	18.14%
TO CONVICT	276	16.3%	220	20.91%
TO CURE	235	20.0%	192	23.44%
TO ENTERTAIN	219	19.63%	209	23.44%
TO EVALUATE	275	13.09%	253	15.42%
TO FIRE	260	18.08%	256	17.97%
TO FRIGHTEN	241	16.6%	207	19.32%
TO HIRE	244	17.21%	198	23.74%
TO INSTRUCT	245	16.33%	248	17.34%
TO INTERROGATE	261	13.41%	236	17.8%
TO INTERVIEW	270	18.15%	231	20.78%
TO INVESTIGATE	271	16.97%	239	16.74%
TO LECTURE	287	14.63%	243	17.28%
TO PUNISH	243	18.93%	227	23.79%
TO RESCUE	218	20.64%	206	22.33%
TO SERVE	271	17.34%	214	20.09%
TO TEACH	243	17.7%	208	21.63%
TO TERRORIZE	279	8.96%	244	15.98%
TO WORSHIP	264	14.02%	249	17.67%

[a] Consistency = the percentage of distinct features produced by two or more workers.

from the main opposition between agent and patient role, some thematic roles for some verbs are clearly better defined than others. For instance, just compare the consistency rate of the agent roles of TO ACCUSE and TO TERRORIZE with those for the verbs TO CURE and TO RESCUE. McRae et al. (1997b) see lack of consistency as a by-product of the fact that those roles can be realized in many possible ways: i.e., the range of people who typically *accuse* or *terrorize* is more varied than those who *cure* or *rescue*.

Gold standards Table 6.2 summarizes the distribution of distinct features for each verb-role pairing in the role-based and role-base-expanded datasets, i.e., in the two collections that will be used as gold standard for the evaluation of our model.

By removing the temporal characterization from our slot-based norms, i.e., by aggregating the features produced for each verb-role pairing, we obtained a total of 7,290 distinct role-based features. Of these, 3,923 were associated with an agent role and 3,367 with a patient role. The difference between the average

Table 6.2 *Distinct features in the gold standard datasets*

	Agent Features		Patient Features	
	Role-based	RB expanded	Role-based	RB expanded
TO ACCUSE	213	1,759	166	1,546
TO ARREST	190	1,481	180	1,677
TO CONVICT	204	1,693	157	1,138
TO CURE	175	1,184	142	1,204
TO ENTERTAIN	176	1,320	156	1,223
TO EVALUATE	226	1,788	187	1,650
TO FIRE	198	1,633	188	1,611
TO FRIGHTEN	193	1,629	150	1,205
TO HIRE	180	1,534	141	1,222
TO INSTRUCT	185	1,488	197	1,710
TO INTERROGATE	196	1,615	182	1,498
TO INTERVIEW	198	1,689	174	1,365
TO INVESTIGATE	208	1,643	173	1,255
TO LECTURE	227	1,912	181	1,633
TO PUNISH	182	1,619	166	1,429
TO RESCUE	155	1,488	150	1,462
TO SERVE	203	1,717	161	1,298
TO TEACH	191	1,586	146	1,116
TO TERRORIZE	224	1,696	182	1,309
TO WORSHIP	199	1,480	188	1,260

number of agent features produced for each verb ($M = 196.15$, $SD = 18.31$) and the average number of patient features ($M = 168.35$, $SD = 17.21$) reaches statistical significance ($t = 7.15$, $df = 19$, $p < 0.001$).

By automatically expanding our features with synonyms available in WordNet, we put together a dataset composed by 59,765 distinct expanded-role-based features, 31,954 for the agent roles and 27,811 for the patient ones. On average, each verb is associated with 1,597.7 agent ($SD = 164.05$) and 1,390.55 patient features ($SD = 194.29$). A paired Student's t-test reveals that such difference is significant ($t = 4.33$, $df = 19$, $p < 0.001$).

6.3 A Distributional Model of Thematic Roles

In computational linguistics, the concept of thematic role is often evoked when referring to two intercorrelated branches of research: the design of lexical resources (e.g., VerbNet, FrameNet, and PropBank), each typically implementing a different idea of what a role is, and the development of tools apt to annotate a sentence with the roles fulfilled by the verbs arguments, given a predefined list of semantic frames or thematic role labels.

Differently from these mainstream approaches, and in line with the work by Reisinger et al. (2015), we adopt a neo-Davidsonian perspective (i.e., we view roles as second-order properties), and we do not see thematic concepts as primitive entities, but as verb-specific concepts represented as clusters of features organized in a prototypical fashion (Dowty, 1991; McRae et al., 1997b). Our assumption in this chapter is that the features entering into the definition of thematic roles depend on the generalized knowledge about the events expressed by verbs. In particular, we argue that important aspects of such knowledge depend on the way verbs are used in linguistic contexts, and that therefore they can be modeled with distributional information automatically extracted from corpora. We are thus dealing with a problem of automatic lexical acquisition, which we tackle in an unsupervised manner, by relying on the minimal possible number of assumptions. Our aim is to present a computational model to extract from corpora the features characterizing verb-specific roles, which we test on the norms presented in Section 6.2. In this section, we review useful insights we borrowed from related literature on distributional semantics (Section 6.3.1) and on the automatic extraction of event chains from corpora (Section 6.3.2), and we present a short description of the core aspects of our model (Section 6.3.3).

6.3.1 Thematic Information in Distributional Semantics

Unsupervised corpus-based models of semantic representation (Sahlgren, 2006; Lenci, 2008; Turney and Pantel, 2010), commonly labeled as vector/semantic/word spaces or distributional semantic models (DSMs), have been established in the last thirty years as a valid alternative to traditional supervised and semisupervised methods. Among the many factors contributing to this success, probably the most cited is the fact that these models are faster and less labor-demanding than manual annotation and semisupervised models.

Another key factor, crucial for the work we present here, is that such models do not need prior knowledge other than that required to implement the so-called Distributional Hypothesis (Harris, 1954; Miller and Charles, 1991). This hypothesis has been received in the NLP literature as a working assumption roughly stating that the similarity of the contexts in which two linguistic expressions occur is a measure of their similarity in meaning (see Sahlgren, 2008, for a more in-depth discussion). This, in turn, is the corollary of another working assumption: that the meaning of a linguistic item is reflected in the way it is used.

Implementation of the distributional hypothesis depends on a few vaguely defined concepts, and the whole literature on DSMs is centered on the characterization of these concepts:

- *Linguistic expressions*: What kind of linguistic expressions can be characterized in distributional terms?
- *Context*: What is the most effective way to characterize the linguistic behavior of our target expressions?
- *Similarity*: How can we compare the linguistics contexts and what kind of semantic similarity can we model?

All existing DSM models incarnate alternative answers to such issues. Restricting this quick summary to DSMs representing words (see Turney and Pantel, 2010, for an overview of the possible target expressions), typically these models are built by scanning a corpus for all occurrences of the target expressions, identifying their contexts, and representing the words by context frequencies in a co-occurrence matrix. Contexts can be windows of words, syntactic relations, patterns of parts of speech, chapters, documents, and so forth (see Sahlgren, 2006, for a comparative review). Generally, the raw co-occurrence matrix is manipulated by (1) weighting the frequencies for highlighting meaningful word-context associations and (2) reducing dimensionality to create dense vectors of latent features for ignoring unwanted variance and/or for computational efficiency reasons. Each vector in the final matrix is assumed to represent the distributional signature of a target word, and is used to calculate the similarity with all the other words of interest according of a chosen vector similarity measure, typically the cosine. (For a critical overview of the commonly adopted technical solutions, see Bullinaria and Levy, 2007, 2012; Lapesa and Evert, 2014.)

Even if the DSM we present in these pages mostly conforms to this general pattern, to the best of our knowledge, no previous system has been proposed to extract the kind of information we are interested in. DSMs have been widely used for the SRL task (e.g., Erk, 2007; Collobert et al., 2011; Zapirain et al., 2013; Hermann et al., 2014; Roth and Lapata, 2015), but mostly to enhance the performance of a SRL classifier, as an ancillary source of information for a task based on a concept of semantic role that is incompatible with the one adopted in these pages.

Our model is directly inspired by works exploiting a distributional space in which linguistic expressions are characterized on the basis of the syntactic environment in which they occur, that is, syntax-based DSMs (e.g., Grefenstette, 1994; Lin, 1998; Padó and Lapata, 2007; Baroni and Lenci, 2010). In these models, syntactic environments are obtained by extracting from shallow-processed or full-parsed text dependency paths such as those linking a verb to its subject or its object. For instance, given the sentence *the supermodel left the catwalk*, in a syntax-based model the distributional entry for the verb TO LEAVE is enriched with a reprocessing of the `dependency:filler` patterns `sbj:supermodel` and `obj:catwalk`. Syntax-based DSMs have proved

to be useful in many semantic tasks. However, the branch of research that uses such DSMs to model thematic fit is the most similar to ours, for two reasons: First, our work and that reviewed in the next subsection share the same view on the usage-based nature of thematic roles; moreover, we all adopt the working assumption that syntactic slots can be seen as rough approximation of semantic roles, at least in a corpus-based model.

The concept of *thematic fit* refers to the appropriateness of a lemma as a filler of a given verb-specific thematic role for a verb. The cognitive relevance of this notion has been widely proved and tested in psycholinguistics (for a review, see McRae and Matsuki, 2009), where thematic fit judgments are typically collected by asking speakers to rate the plausibility of a lemma being a filler of a given thematic role for a given verb. Such a notion is intimately related to, although not equivalent to, the notion of selectional preference, the main difference being the nature of the involved elements: discrete semantic types in the case of selectional preferences, gradient compatibility of an argument with a thematic role in the case of thematic fit.

To the best of our knowledge, Erk et al. (2010) were the first to evaluate a syntax-based DSM against human-generated thematic fit judgments. In the exemplar model described by these scholars, i.e., the EPP model first introduced by Erk (2007), plausibility scores for argument filler are computed by measuring the similarity of the new candidates with all the previously attested fillers for that verb-role pairing. Crucially, the distributional knowledge extracted in this model comes from two corpora, or from different uses of the same corpus: a "primary" corpus, used to obtain information about verb-argument co-occurrences, and a "generalization" corpus, exploited to extract similarity measures between argument fillers. Erk and colleagues tested their proposal by correlating the plausibility values produced by the system against the human-generated judgments collected by McRae et al. (1998) and those by Padó (2007). The crucially different sparsity degrees of the stimuli in the two datasets clearly affected the performance of the model, which, all things considered, mildly correlated with human judgments only on the latter dataset.

Similar results, with slightly higher correlations, were reported by Baroni and Lenci (2010) when evaluating their framework, Distributional Memory (DM), against the same judgments. In contrast to the practice of developing different DSMs for different tasks, DM is a framework in which co-occurrence information is extracted just once and represented into a third-order tensor that functions as a semantic knowledge repository. When tackling a specific task, the DM tensor is then manipulated to create the task-specific DSM as needed, without resorting back to the corpus. In modeling thematic fit, Baroni and Lenci (2010) showed how their tensor can be manipulated to derive a matrix in which the vectors are the target lemmas and the dimensions are `dependency:filler` patterns. This syntactic DSM, analogous to the representation exploited by Erk

et al. (2010), is then used to identify, for each verb, its typical subject and object fillers, to built their centroids (i.e., their "prototypical" vectors), and to predict thematic fit for a given noun-role-verb by measuring the distance between the target noun and the verb-role centroid.

Greenberg et al. (2015) compared the performance of the model by Baroni and Lenci (2010) with those that can be obtained by two different role-based DSMs. Moreover, they experimented with different methods to calculate the prototypical vector set for each verb-role. The results of this comparative work, evaluated against the datasets by McRae et al. (1998) and by Padó (2007), together with the instrument and location roles judgments by Ferretti et al. (2001), showed a slight superior performance for the DM-based model,[5] and a clear constant improvement in using agglomerative clustering to build the prototypical filler of a verb-specific role. Finally, Lenci (2011) goes further in the investigation of the thematic fit phenomenon by using the same DM-derived matrix as Baroni and Lenci (2010) to model how argument expectations are updated on the basis of the realization of the other roles in the verbs argument structure. Evaluated against data from Bicknell et al. (2010), the best settings of this model obtained a 73-84% hit accuracy rate.

Another strand of research that has been inspirational for our work includes those works that try to model feature norms information for concrete concepts by means of a DSM. The first attempts to automatically extract short descriptions of this sort are described in Almuhareb and Poesio (2004, 2005) and Barbu (2008). These approaches were quite limited in their scope, being focused on a restricted set of semantic relations, two in the former studies, six in the latter. To the best of our knowledge, Baroni et al. (2010) were the first to tackle an unconditional version of this task. Their model Strudel extracts properties by looking at the distribution of superficial patterns like [Concept]_*is*_*ADV*_[Property] (as in *the grass is really green*) or [Property]_*of*_[Concept] (as in *pack of wolves*). The key intuition is that a strong semantic link between a concept and a property reflects in their co-occurrence in a great variety of different patterns. Evaluated against the ESSLLI dataset, the authors reported a precision score of 23.9%, to date the highest score registered for the unconstrained extraction of feature-like ⟨*concept, property*⟩ pairs. As argued by Devereux et al. (2009), a major limitation of the Strudel approach is that the semantic relations between concepts and properties are characterized only implicitly, i.e., by means of superficial patterns. In fact, Strudel can

[5] For the sake of completeness, it should be noted that Erk et al. (2010) also compared the results obtained by exploiting, as a primary corpus, a role-semantic rather than a syntactic annotation, and report a slight advantage of the former over the latter. As noted by the authors, however, the presence of the many sources of variance (manual vs. automatic annotation, corpus size) doesn't allow any firm conclusion from these results.

be seen as an unconstrained model to extract feature-like ⟨*concept, property*⟩ pairs.

Devereux, Kelly, and colleagues (Devereux et al., 2009; Kelly et al., 2010) were the first scholars to try to automatically extract feature-like ⟨*concept, relation, property*⟩ triples. They tried to identify the prototypical properties of a concept and to explicitly characterize the type of their relation. The model they proposed articulates in two phases: first, manually generated syntax-based rules were used to extract a set of candidate ⟨*concept, relation, property*⟩ triples; then these triples were ranked on the basis of the conditional probabilities of concept and feature classes derived from the McRae dataset. As reported by Kelly et al. (2010), when evaluated against the ESSLLI dataset, their best model obtained a precision score of 19.43% for the identification of ⟨*concept, property*⟩ pairs and 11.02% when looking for ⟨*concept, relation, property*⟩ triples.

Kelly et al. (2013) moves on by proposing a model that exploits syntactic, semantic, and encyclopedic information. This model starts by applying a series of rules to extract meaningful paths from the syntactic annotation available in two corpora: an encyclopedic corpus and a general corpus. Then the model weights each candidate triple first by using a linear combination of four metrics and later applying the same reweighting strategy as in Devereux et al., 2009; Kelly et al., 2010. When evaluated against the same settings used by Baroni et al. (2010), their best models obtain a precision score of 13.39% for the identification of ⟨*concept, property*⟩ pairs and 5.02% when looking for ⟨*concept, relation, property*⟩ triples.

6.3.2 A Wider Context: Narrative Event Chains

Another branch of research investigating an issue related to ours focuses on the unsupervised characterization of *Narrative Event Chains*, defined as partially ordered set of events involving the same protagonist (Chambers and Jurafsky, 2008), where an event is represented by a verb together with its arguments. The following example, adapted from Chambers and Jurafsky (2009), describes a chain in which the protagonist is being prosecuted. The sequence of the events in this chains can be summarized as: the protagonist admits something and pleads (guilty), before being convicted and sentenced. Formally, this chain can be represented as a tuple (L, O), where L is a set of ⟨*event, argument slot*⟩ tuples and O is a partial temporal ordering:

L = ⟨*admit, subject*⟩, ⟨*plead, subject*⟩, ⟨*convict, object*⟩, ⟨*sentence, object*⟩

O = {(*plead, convict*), (*convict, sentence*), ...}

The unsupervised characterizations of event chains and related issues, such as the induction of event schemas and the temporal ordering of events, have

been tackled by relying on different source data, e.g., text corpora (e.g., Chambers and Jurafsky, 2008, 2009; Chambers, 2013; Balasubramanian et al., 2013) vs. crowdsourced descriptions (e.g., Regneri et al., 2010; Frermann et al., 2014), and on different families of approaches, e.g., graph-based methods (e.g., Regneri et al., 2010; Balasubramanian et al., 2013), probabilistic approaches (e.g., Cheung et al., 2013; Chambers, 2013; Frermann et al., 2014) or distributional learning (e.g., Chambers and Jurafsky, 2008, 2009).

There is a close relationship between the concept of event chain and the entailment-based concept of semantic role we adopt in this chapter. In a sense, part of the verb-specific entailments we aim to model is what happens to a protagonist (i.e., the role filler) in a prototypical event chain if we take our target verb as a reference point. As an example, let us go back to the prosecution narrative chains mentioned earlier and suppose that we have proved that they describe a prototypical sequence of events. At least some of the entailments associated to the patient of the verb TO CONVICT correspond to what happens to her/him before, during, and after the conviction event takes place. These entailed actions and properties may be found among the events that compose a prototypical narrative schema containing our target ⟨*event, argument*⟩ pairing: for instance, *s/he admits, s/he pleads (guilty)*, and *s/he is convicted*.

With this parallelism in mind, we looked at the seminal models by Chambers and Jurafsky (2008, 2009) for useful insights and intuition to integrate into our model, especially in light of the methodological affinities between our works. The starting point of Chambers and Jurafsky (2008) is the "narrative coherence" assumption: verbs whose arguments belong to the same coreference chain are semantically connected, and more likely to participate in a narrative chain. Briefly, the model proposed by these authors articulates in three steps. In the first step, the protagonist and the subevents are identified by first calculating the strength of association between pairs events, where the association score is a function of how often two events have a coreferring entity, and combining these pairwise associations into a global narrative score. Evaluated with a variation of the cloze task (Taylor, 1953), such a method shows a 36% improvement over baseline. Association scores are later fed to an agglomerative clustering algorithm to construct discrete narrative chains. In parallel, a two-stage machine learning architecture is used to temporally order these connected subevents, obtaining a 25% increase over a baseline for temporal coherence.

Chambers and Jurafsky (2009) extended these results by dealing with two limitations of their previous proposal: the lack of information concerning the role or type of the protagonist and the fact that only one participant was represented. As a solution to the former issue, the authors propose the notion of "typed" narrative chains, that is, an extension of the notion of chain in which the argument shared between events is defined by being a member of a given set

of lexical units, nouns cluster, or other semantically motivated group. The second extension results in the introduction of the concept of "narrative schema," that is, an extension of the notion of narrative chain that models the entire narrative of the document by generalizing over all the actors involved in a set of events. When tested against the same dataset of Chambers and Jurafsky (2008), the joint effect of both extensions resulted in a 10% increment over the performance of the previous model.

6.3.3 The Core of a Neo-Davidsonian DSM for Semantic Roles

In the rest of this section we describe the core characteristics of a DSM incorporating a neo-Davidsonian view of verb-specific roles as clusters of prototypical features derived from corpus-based distributional data. In the next section we describe how we translated this model into an algorithm that we tested against the human-elicited propertieds described in Section 6.2.

Our main assumption is that (at least a subset of) the entailments associated with the specific roles of a target verb derive from the actions and properties associated with the role fillers in prototypical narrative schemata containing our target verb. Given a verb v and its specific role r_v, we define $f_1, \ldots, f_n$ as the n-most prototypical noun fillers of r_v: for instance, if r_v is the patient role of TO CONVICT, the fillers can be *defendant*, *prisoner*, etc. Let $s_1, \ldots, s_n$ be the narrative sequences of events in which the role-filler pairs $\langle r_v, f_i \rangle$ occur in a corpus. Each sequence s_i can be regarded as a broader scenario including the event expressed by the target verb v and the filler f_i for the role r_v. We then provide the following distributional characterization of verb-specific thematic roles:

> The verb specific role r_v is the set of the predicates most associated with its fillers $f_1, \ldots, f_n$ in the narrative sequences $s_1, \ldots, s_n$.

This framework thus relies on insights derived from both strands of research outlined in this section: from Erk et al. (2010) and subsequent works we borrowed the idea that thematic fit can be modeled by means of a syntax-based DSM; from Chambers and Jurafsky (2008) and subsequent works we borrowed the idea that the discourse structure, and in particular coreference chains, can be used to model sequences of events belonging to larger scripts or scenarios. In the final model, the semantic content of each verb-specific thematic role is represented by the set of predicates that meet the following two conditions:

- they are strongly associated with the prototypical fillers $f_1, \ldots, f_n$ of a verb-specific role r_v;
- one of its argument frequently belongs to the same coreference chain as the filler of r_v.

These sets of entailments are identified by combining two contextual representations: a distributional syntax-based representation and a coreference-based representation. The distributional contextual representation is built in a three-step process:

1. A dependency extraction phase, during which a corpus is scanned to identify and manipulate all the relevant syntactic relations headed by a verb v. As noted by other authors (e.g., Preiss et al., 2007), such a process should be carefully tuned on the behavior of the specific parser used to annotate the input corpus.
2. Syntax-based co-occurrency frequencies are then used to calculate the association score between each verb-slot pairing and its fillers $f_1, \ldots, f_n$. Given the symmetrical nature of association measures, this information can be used to model both direct and inverse selectional preferences (Erk et al., 2010). Accordingly, this step is used to select, for each verb-specific role r_v, its prototypical fillers as well as, for each filler, the prototypical verb-specific role in which it occurs. In what follows, we use the notation `relation-1:predicate` to refer to a construction representing both the inverse relation linking a lemma to its head, as well as the head. Its intuitive meaning can be paraphrased as "(the filler) is the *relation* of *predicate*," e.g., `obj-1:eat` indicates a filler (e.g., *apple*) is the object argument of the verb TO EAT.
3. Finally, direct and inverse preferences are manipulated to associate each target r_v with a set of contextual `relation:predicate` constructions, obtained by interpreting the inverse selectional preferences of each r_v prototypical filler as clues of semantic relatedness. As an example, let us suppose that the target r_v is the patient role of the verb TO WRITE and that in the previous step we learned that its top associated fillers are *letter* and *book*. Let us assume that these nouns are strongly associated with the object position of {TO RECEIVE, TO SEND, TO COMPOSE} and {TO READ, TO PUBLISH, TO DEDICATE}. In this step we would elaborate on this picture to identify a set of candidate entailments such as the one between the patient of TO WRITE and `obj-1:publish` (i.e., what is written is typically published), or TO WRITE and `obj-1:read` (i.e., what is written is typically read).[6]

The distributional contextual representation collects events and properties that are related to our target verbs, but only a part of these are entailment patterns that may reasonably be assumed to enter into the definition of verb-specific thematic roles. For instance, while it is fairly plausible to presume the

[6] Note that we are focusing solely on the object position of a verb just for the sake of exposition. As will become clear in the following section, this line of reasoning applies to all semantic roles we may find useful.

existence of an entailment relation between TO WRITE and TO PUBLISH, the relation between TO WRITE and TO COMPOSE is clearly one of near-synonymy. In fact, a crucial assumption of our model is that the distributional features characterizing verb roles belong to the event sequences including the target verb and its fillers. This is indeed the case of TO WRITE and TO PUBLISH, which can be assumed to be part of a larger book production scenario. We identify sequences of events including both the target verb and the extracted distributional context information with the following procedure of coreference-based contextual representation:

1. We extract from a coreference-annotated and parsed corpus all the verbs and nominal predicates whose argument typically belongs to the same coreference chains of the fillers $f_1, \ldots, f_n$ of our target verbs. Crucially, in this passage we keep track of the syntactic relation between each verb and the entities involved in each coreference chain. For instance, the text *I wrote you a note the other day. Did you read it? Yes, and I posted it online* contains chains linking the object of our target verb TO WRITE, the object of the verb TO READ, and the object of the verb TO POST.
2. From each coreference chain, we extracted, for each target verb-specific role r_v (e.g., the object role of the verb TO WRITE), all the inverse dependencies involving each of the entities that corefer with our target verbs filler. In our example, this means isolating the contextual constructions `obj-1:read` and `obj-1:post`, jointly meaning something like "(the filler of our target verb-specific role corefers with) the object of the verb TO READ and the object of the verb TO POST."
3. For each r_v, we removed the inverse dependencies missing from the distributional contextual representation and use the filtered coreference-based co-occurrence frequencies to calculate the strength of association between each target r_v (e.g., the object role of the verb TO WRITE) and each contextual construction (`obj-1:read` and `obj-1:post`). The most associated constructions are precisely the distributional features we use to characterize the entailments associated with r_v.

6.4 Experiments with Our Neo-Davidsonian Model

We tested the validity of our approach by evaluating how many of the speaker-generated entailment patterns collected in the experiment described in Section 6.2 we are able to automatically extract from an annotated corpus. To test the relative strength of the different sources of information, three different DSMs were created: the full model, implementing all the passages described in Section 6.3.3; a coreference model, in which only coreference-based

information were used; a distributional model, based solely on distributional contextual representations. We dubbed the latter two models *quasi-Davidsonian*.

All models were trained on a coreference-annotated and parsed version of the British National Corpus[7] (BNC; Aston and Burnard, 1998), a 100M words corpus of British English language productions from a wide range of written (90%) and spoken (10%) sources, built in the first half of the 1990s. The corpus has previously been POS-tagged and lemmatized with the TreeTagger[8] (Schmid, 1994), parsed with MaltParser[9] (Nivre et al., 2007), and coreference-annotated with BART[10] (Versley et al., 2008).

In collecting our gold standard role descriptions, we followed the settings of McRae et al. (1997b) which focused solely on the agent and patient proto-roles. Consequently, the following experiments address only these two roles. To limit data sparsity, we included in our test set only those verbs of McRae's list that occurred in the BNC more than 1,000 times, a condition that was not met by three verbs: TO TERRORISE ($f = 115$), TO INTERROGATE ($f = 274$), and TO WORSHIP ($f = 369$).

6.4.1 The Full Neo-Davidsonian Model

Implementation of the full model follows the general picture outlined in Section 6.3.3. As a first step, we scanned the syntactic annotation of the corpus and applied a set of parser-specific rules to handle such problematic phenomena as conversion of the passive diathesis, identification of the antecedents of relative pronouns, treatment of conjunct and disjunct; and identification and treatment of complex quantifiers such as "a lot of."

We then looked in the corpus for instances of relevant syntactic relations, and for each occurrence we identified the element heading the relation and the lemma filling the argument position, thus obtaining a tuple of the form: ⟨*verb, relation, filler*⟩. In these experiments, the set of dependency relations we are interested in is composed by:

- `sbj`: *the professor wrote the letter* → ⟨*write, sbj, professor*⟩;
- `obj`: *the professor wrote the letter* → ⟨*write, obj, letter*⟩;
- `prd`: *the letter became famous* → ⟨*letter, pred, famous*⟩.

Following Erk et al. (2010), we see the `sbj` and `obj` relations as surface approximations of the agent and patient proto-roles. As such, they will be used both for characterizing the selectional preferences of our target verbs, as well as the inverse selectional preferences of their prototypical fillers. The `prd` relation,

[7] www.natcorp.ox.ac.uk [8] www.cis.unit-muenchen.de/∼schmid/tools/TreeTagger/
[9] www.maltparser.org [10] www.bart-coref.org

on the other hand is only used to extract the inverse selectional preferences of the prototypical fillers of our target verbs. This way, we extract all those properties that are typically described by adjectives or nouns.

We used the frequency of each ⟨*verb, relation, filler*⟩ tuple to calculate the strength of association between verb-specific roles (i.e., ⟨*verb, relation*⟩ pairs) and fillers. In our experiments we used positive Local Mutual Information (pLMI) to calculate the strength of association between a target entity (e.g. a verb-specific role r_v) and a given context (e.g. a contextual construction). Local Mutual Information (LMI; Evert, 2009) is defined as the log ratio between the join probability of a target t_i and a context c_j and their marginal probabilities, multiplied by their joint frequency:

$$LMI(t_i, c_j) = f(t_i, c_j) * \log_2 \frac{p(t_i, c_j)}{p(t_i) * p(c_j)} \tag{6.1}$$

LMI is a version of the Pointwise Mutual Information (PMI; Church and Hanks, 1991) between a target and a context weighted by their joint frequency, usually preferred to PMI to avoid its characteristic bias toward low-frequency events. Positive LMI is obtained by replacing all negative values with 0:

$$pLMI(t_i, c_j) = max(0, LMI(t_i, c_j)) \tag{6.2}$$

We used these statistics to select:

- *Direct selectional preferences*: The top 50 fillers associated with each target ⟨*verb, relation*⟩ tuple, where the relation can be either `sbj` or `obj`. For each target verb, therefore, we collected 50 subject and 50 object fillers;
- *Inverse selectional preferences*: The top 100 ⟨*verb, relation*⟩ tuples associated with each filler, where the relation can be either `sbj`, `obj`, or `prd`.

The distributional-based contextual representation is built by associating each target verb-specific roles r_v with the top ⟨*verb, relation*⟩ tuples of their top fillers. Alternatively said, the output of this phase will be obtained by merging, for each r_v, the inverse preferences of its prototypical fillers, thus obtaining a set of `relation-1:verb` contextual constructions.

In a second phase, for each r_v, we parsed all the coreference chains involving its fillers and isolated the verbal head or predicate of each entity in a ±2-sentences window that belongs to the chain of our verbs filler. We chose to focus on a portion of the coreference chain centered on the target verb to avoid those events of the narrative chains that are not directly related to our target event. We leave to future investigation the evaluation of the effect of this hyperparameter. In this passage, we track the dependency relations between the coreferring entities and their heads, thus obtaining sets of `relation-1:verb` contextual constructions analogous to the ones exploited as contexts in the distributional representation.

The two sets of contexts, i.e., the one used in the distributional model and the one used in the coreference model, are indeed comparable notwithstanding their different natures. They both encode different kinds of semantic relatedness: relatedness due to the sharing of the same sets of fillers in the case of the distributional contexts; relatedness due to the participation to the same event chains in the case of the coreference contexts. We take advantage of this compatibility by filtering the latter on the basis of the former. That is, for each verb-specific role r_v, we retain only the `relation-1:verb` contextual constructions that are shared between the distributional and the coreference representations. Finally, we resort to the coreference chains to extract the co-occurrence frequency between the r_v and the selected contextual constructions, and to calculate their association with pLMI. In our view, these top associated contextual constructions provide a distributional representation of the entailment patterns licensed by our verb-specific roles.

Table 6.3 in Appendix 1 reports the top associated contextual constructions that our model extracted for the agent and patient roles of the verbs TO ARREST and TO PUNISH. Intuitively, the high association between the agent role of the verb TO ARREST and the contextual constructions `sbj-1:hold` and `sbj-1:imprison` could be paraphrased as *s/he who arrests someone also holds him/her/someone else* and *s/he who arrests someone also imprisons him/her/someone else*. On the other hand, the high association between the patient role of the verb TO PUNISH and the contextual constructions `obj-1:torture` and `sbj-1:desperate` could be paraphrased as *s/he who is punished may be also tortured* and *s/he who is punished may be desperate*.

6.4.2 Quasi-Davidsonian Models

To evaluate the relative importance of the two souls of our neo-Davidsonian DSM, i.e., the distributional and the coreference-based components, we created two different DSMs, each modeling exclusively one kind of information. In the distributional model, the association between r_v and the contextual constructions is calculated solely on distributional basis, without trying to account for the patterns in the narrative structures that can be extracted from the coreference annotation. A coreference-based model relies solely on the information that can be extracted from the coreference chains, without resorting to syntax-based distributions to filter out infrequent fillers.

6.4.3 Evaluation

DSMs are usually evaluated extensionally, that is, by recording their performance on tasks that are supposed to tackle some crucial aspects of the human semantic memory. Typical tasks are to mimic the intended behavior of the

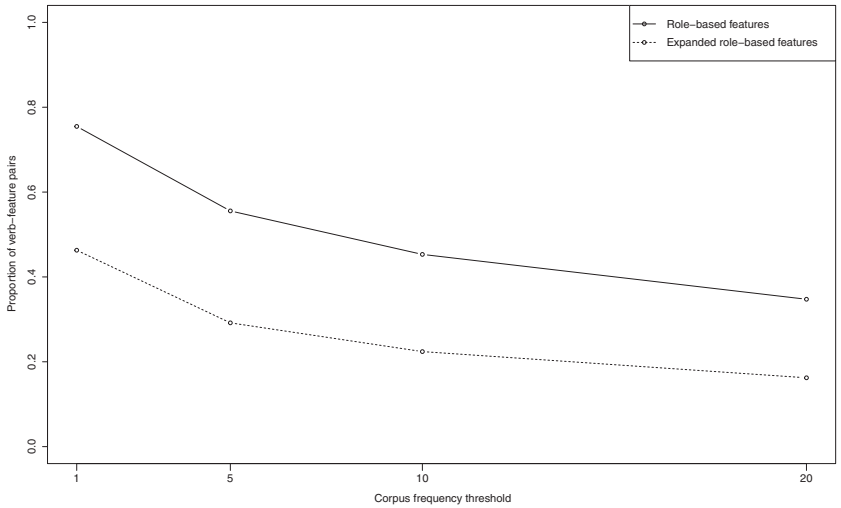

Figure 6.2 Proportion of verb-feature pairs in a $\pm$2-sentences window, modulated by application of different frequency thresholds.

proficient speaker in tests like the synonym test questions from the Test of English as a Foreign Language the correlation with human-generated linguistic metajudgments, and clustering words to emulate some available semantic classification or how well they serve as features for machine learning algorithms.

Intensional methods, in which the validity of the semantic knowledge encoded in a DSM is directly assessed, are less common. Practices of this sort include the manual (usually crowdsourced) evaluation of the nearest neighbors returned by DSMs for target items, or the test against a prepared dataset of valid target-context associations such as speaker-elicited features.

In these pages we adhere to this latter tradition and evaluated our model against the two gold standard datasets described in Section 6.2.3, i.e., the role-based dataset obtained by stripping the temporal characterization from the descriptions collected in Section 6.2, and the expanded-role-based datasets built by enriching the role-based features with synonymic information available in WordNet.

Before moving to the evaluation of our model, however, it is wise to assess whether our training corpus, the BNC, actually contains the kinds of information collected in our gold standards. An easy way to do so is to count the proportion of human-generated features that co-occur with the target verbs within a given windows size, thus adapting the paradigm exploited by Schulte im Walde and Melinger (2008) to investigate verb semantic associations. Consistent with the settings of our DSMs, we fixed our window size to $\pm$2 sentences

and excluded from our analysis the three verbs whose absolute frequency was below the 1,000 occurrences threshold (i.e., TO TERRORIZE, TO INTERROGATE, and TO WORSHIP).

Figure 6.2 shows the proportion of verb-feature pairs from the role-based dataset (*solid line*) and from the expanded role-based dataset (*dotted line*) that co-occur in the BNC with a minimum frequency of 1, 5, 10, and 20 (x-axis). Focusing on the most appropriate threshold given the corpus size, i.e., a minimum frequency of 5, we can see that a bit more than half of the verb-feature pairs (55.56%) from the extended role-based dataset can be traced in the BNC, and this proportion decreases to less than one third if we look for the verb-feature pairs from the role-based dataset (29.18%). These numbers seem to confirm the shared belief that there are crucial differences in the information that can be extracted from corpora and the information extracted from human-elicited descriptions. Whereas some authors see corpora-derived measures as "a form of crowd-based measures, where the crowd consists of writers freely creating text on different topics" (Keuleers and Balota, 2015, p. 463), others stress the fact that corpora seem to lack many of the nonlinguistic mental properties available in the norms collections (De Deyne et al., 2015) or the fact that norms tend to represent distinctive properties of concepts, whereas texts in corpora report properties that are relevant for their communicative purposes (McRae et al., 2005b).

What is crucial for the present work, however, is the awareness that our models should not try to reach the maximum recall, but rather focus on precision. That is, a model's performance depends on its ability to associate each verb-specific role with features that are attested in our gold standards, notwithstanding its ability to extract *all* the information available in our datasets. This is reminiscent of what happens in many Information Retrieval studies, particularly those involving web search (Manning et al., 2008), which measure the precision of the top k retrieved results. Similarly, we derive the "Precision at k" metric by counting how many features, for each verb-specific role r_v, are attested in the gold standard.

However, the gold standard features are not directly comparable with the contextual constructions `relation-1:verb` extracted by our DSMs. We therefore simplified the latter by removing the specification of the inverse syntactic relation. This way, we are not able to distinguish constructions such as `subj-1:imprison` (*s/he imprisons*) and `obj-1:imprison` (*s/he is imprisoned*). Both contextual constructions are thus conflated into one feature: `imprison`.

To determine whether both the full model and the quasi-Davidsonian model performed better than chance, we implemented a random baseline for each model by replacing every feature with a common noun, verb, adjective, or adverb in the same frequency range. For expediency, we won't report here the

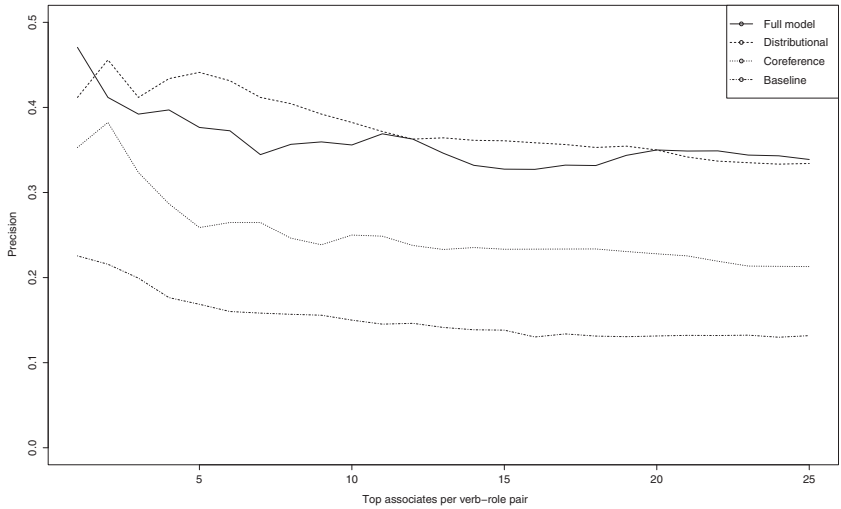

Figure 6.3 Precision of the different models evaluated against the dataset of extended role-based features.

precision of each randomized model, but we do average between them and refer to this baseline as the one obtained from the random model.

Figure 6.3 shows the precision values of the different models for different top k-selected features per verb-specific role (x-axis), evaluated against the extended role-based features. Exact values for reference values of k are reported in Table 6.4 in Section Appendix 2. Results appear to be higher than those reported by the literature on the automatic extraction of feature-like descriptions of concrete concepts (Baroni et al., 2008; Baroni and Lenci, 2010; Kelly et al., 2013), but their magnitude should be better interpreted as another confirmation of the difficulty of the task.

The best-performing models are the full model (*solid line*) and the distributional model (*dashed line*), both performing better than the coreference-based one (*dotted line*). All DSMs, moreover, performed better than the chance level (*dash-dotted line*), whose precision is around 0.15.

A similar pattern, although with lower precision scores, is obtained by evaluating the models against the role-based features, as shown by Figure 6.4 (see scores on Table 6.5). Again, all models perform better than the random baseline (precision ≈ 0.04). Again, the full model and the distributional model registered better scores than the coreference-based model.

There are, however, two main reasons why we would not take this as strong evidence against the utility of coreference-based information in modeling semantic role inferences. First of all, given the preliminary nature of our work,

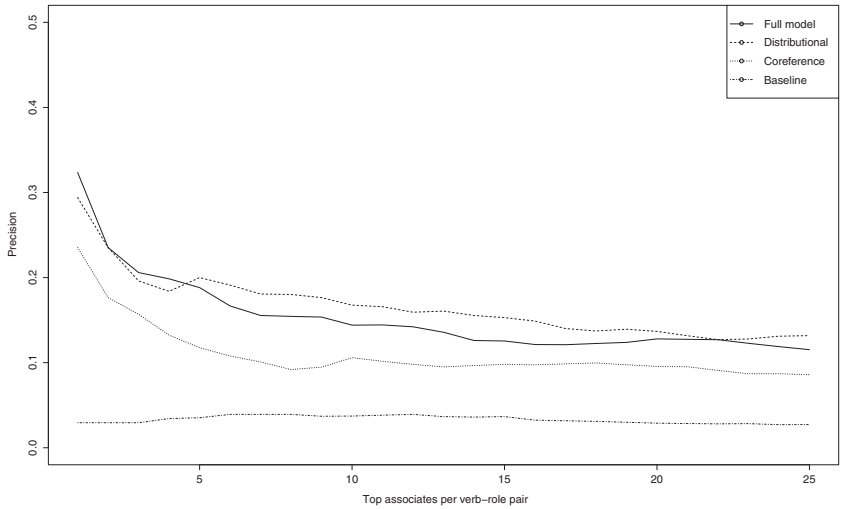

Figure 6.4 Precision of the different models evaluated against the dataset of role-based features.

we did not experiment with many of the settings that could influence the performance of both the full model and the coreference model, including the size of the context. Moreover, there is probably a joint effect of the general difficulty of the coreference annotation task (Recasens et al., 2010; Pradhan et al., 2012) and of data sparsity, due to both corpus size and the neglect of dialogue-related phenomena like implicit arguments (e.g., Ruppenhofer et al., 2011; Roth and Frank, 2013). On the other hand, the performance of the full model itself is a clue in favor of our caution. This model is basically a coreference DSM that exploits the distributional information merely to filter out unwanted features. As a consequence, it is fairly possible that the gap between the full model and the coreference DSM is due to noise that can be eliminated by using a wider context for the coreference chains, by exploiting a larger corpus, or by manually checking the relevant coreference data.

Taken together, these results appear encouraging to us, especially in light of the several limitations in the implementation of our models. Clearly, these weak spots leave plenty of room for future improvements. First of all, we overtly decided to ignore all the properties that can be inferred from dependencies headed by the fillers (in fact we used only inverse dependencies) or from superficial patterns. This will require an in-depth evaluation of the possible strategies to extract this additional information and to integrate it in our model. Moreover, it is possible that the settings we chose for our hyperparameters (e.g., number of

top fillers, association measure) are not optimal for our task. As far as the distributional space is concerned, we drew from previous experience and available comparative works (e.g., Bullinaria and Levy, 2007, 2012; Lapesa and Evert, 2014). The situation has been quite different for the use of coreference information. There was no available comparative literature, and we made our choices mainly drawing from intuition and qualitative analysis of several rounds of preliminary testing.

Finally, it is well known that the evaluation methods we chose underestimate precision. The exemplar contextual constructions in Table 6.3 from Appendix 1 illustrate this point. In this table, the constructions whose fillers are associated with the target verb-specific role in the gold standard are marked in the *match* column: two check marks for role-fillers pairings that are attested in both gold standards, one check mark for those pairings that are attested only in the extended role-based norms. Even a quick look at the unmarked features associated with the verb TO ARREST shows a high number of false negatives: *s/he who arrests may even release*, *s/he who arrests may detain*; *s/he who is arrested may bail*; *s/he who is arrested may be oppressed*; *s/he who is arrested may have been recaptured*; *s/he who is arrested may be proclaimed* (e.g., innocent); *s/he who is arrested may be inhibited*; *s/he who is arrested may have abducted someone*. Arguably, we could have chosen a less conservative evaluation method, such as a feature verification paradigm. Scholars working on the automatic extraction of concrete concepts features report increases in precision as high as 0.4 when switching from a norm-based evaluation to an evaluation based on speakers' judgments (Kelly et al., 2013). Crowdsourcing techniques analogous with those developed by Reisinger et al. (2015) easily can be adapted for our purposes. However, such a choice would have come at the price of a higher number of false positives, mainly because it is often possible to find a context in which a role-feature pair may be true, even if this association is not particularly meaningful. Once again, we opted for the conservative choice, thus leaving the use of different evaluation techniques to future investigations.

In closing, it is worthwhile to stress that another consequence of the preliminary nature of our work has been the choice to restrict our target verbs to those investigated by McRae et al. (1997b). It is our opinion that the generalization of our results to other verbs would require control of many random and fixed effects, including several shades of ambiguity (e.g., lexical ambiguity, syntactic ambiguity), sociolinguistic issues (e.g., corpus data could be biased toward less prototypical uses of a verb), even theoretical considerations (some classes of verbs [e.g., light verbs] are probably harder to characterize, automatically or manually). Owing to space limitations, however, we must leave investigation of this crucial issue to future works.

6.5 Conclusion

This paper has introduced a novel unsupervised method to characterize the semantic content of verb-specific agent and patient proto-roles as bundles of presuppositions and entailment relations. Our primary intent was to test whether and to what extent semantic knowledge automatically extracted from text can be used to infer the kinds of entailments on which semantic roles are grounded. At the same time, by tackling this issue we implicitly provided evidence in favor of the idea that at least part of the knowledge about events manifests itself in the way verbs are used in a communicative environment, and that part of this generalized knowledge can be distilled from the linguistic productions available in corpus. In the view adopted in these pages, which we borrowed from Dowty (1991) and McRae et al. (1997b), it is exactly this kind of knowledge that works as a source from which the semantic content of thematic roles, by a sort of clustering process, is carved.

We evaluated different implementations of our method against a dataset of human-elicited descriptions collected with a modified version of the McRae paradigm (McRae et al., 1997b) and expanded with lexical knowledge from WordNet. In each setting, all of our models performed well above the chance level. The best-performing models were a purely syntax-based DSM and a coreference-based DSM enhanced by a syntax-based representation, both achieving a precision score between 0.35 and 0.45. Both the behavioral data and the automatically extracted verb-specific properties are freely available for downloading at http://colinglab.humnet.unipi.it/resources/.

The main contribution of our work, however, is not the model itself, but the demonstration that state-of-the-art computational techniques can be easily adapted to reach a decompositional description of the semantic content of thematic roles. To the best of our knowledge, the only related work in the computational linguistics literature is the one by Reisinger et al. (2015). As a consequence, we cannot but speculate over the potential applications that can benefit from our shift in perspective. However, one specific branch of research pops up immediately, i.e., the one focusing on the extraction and representation of Semantic Roles. No decompositional approach available today has the maturity to be used as a complete and usable theoretical framework, and that's probably why we're still stacked with the traditional *I-can't-define-it-but-I-know-it-when-I-see-it* stance on thematic roles, using Dowty's words (Dowty, 1989). However, the theoretical perplexities that drove the theoretical linguists to treat the atomistic view of semantic roles as an inadequate representation of the reality are strictly related to the difficulties that all researchers deal with when working with thematic roles: What is their inventory? How can they be identified? On the basis of which properties? How are they realized in the syntactic structure? The model we proposed in these pages should be seen as an attempt to

look at all these theoretical and practical issue from a different, decompositional, perspective.

Acknowledgments

The authors thank Gaia Bonucelli for taking care of the normalization phase reported in Section 6.2.2. This research received financial support from the CombiNet project (PRIN 2010-2011: *Word Combinations in Italian: theoretical and descriptive analysis, computational models, lexicographic layout and creation of a dictionary*, grant n. 20105B3HE8) funded by the Italian Ministry of Education, University and Research (MIUR).

Appendix

Appendix 1 Exemplar Features for the Verbs "to arrest" and "to punish"

Table 6.3 *Top 10 associated features per role extracted with the full model*

TO ARREST			TO PUNISH		
Role	Feature	Match[a]	Role	Feature	Match
agent	sbj-1:hold	✓	agent	sbj-1:reward	
agent	sbj-1:release		agent	sbj-1:forgive	
agent	sbj-1:charge	✓	agent	sbj-1:catch	✓
agent	sbj-1:say	✓	agent	sbj-1:deserve	
agent	sbj-1:imprison	✓✓	agent	sbj-1:doom	
agent	sbj-1:detain		agent	sbj-1:condemn	
agent	sbj-1:sentence		agent	sbj-1:compound	
agent	sbj-1:remand	✓	agent	sbj-1:tolerate	✓
agent	sbj-1:live	✓	agent	sbj-1:forfeit	
agent	sbj-1:fall		agent	sbj-1:deter	
patient	sbj-1:intern		patient	obj-1:reward	
patient	sbj-1:bail		patient	obj-1:torture	✓
patient	obj-1:oppress		patient	obj-1:lock	
patient	obj-1:recapture		patient	obj-1:unnerve	✓✓
patient	sbj-1:defy	✓	patient	obj-1:whip	
patient	sbj-1:confine	✓	patient	obj-1:humiliate	✓✓
patient	obj-1:proclaim		patient	obj-1:indulge	
patient	obj-1:inhibit		patient	obj-1:torment	✓✓
patient	sbj-1:abduct		patient	sbj-1:desperate	
patient	sbj-1:caution		patient	obj-1:elevate	

[a]*Match:* whether the triple ⟨*verb, role, feature lemma*⟩ is present in the role-based norms (✓✓), in the expanded-role-based norms (✓) or in none of the gold-standard datasets (empty cell).

Appendix 2 Precision at **k** *of the Different Models Evaluated against the Feature-based Gold Standards*

Table 6.4 *Evaluation against the dataset of extended role-based features*

k	Full model	Distributional	Coreference	Baseline
5	0.38	0.44	0.26	0.17
10	0.36	0.38	0.25	0.15
15	0.33	0.36	0.23	0.14
20	0.35	0.35	0.23	0.13
25	0.34	0.33	0.21	0.13

Table 6.5 *Evaluation against the dataset of role-based features*

k	Full model	Distributional	Coreference	Baseline
5	0.19	0.20	0.12	0.04
10	0.14	0.17	0.11	0.04
15	0.13	0.15	0.10	0.04
20	0.13	0.14	0.10	0.03
25	0.12	0.13	0.09	0.03

References

Almuhareb, Abdulrahman, and Poesio, Massimo. 2004. Attribute-Based and Value-Based Clustering: An Evaluation. Pages 158–165 of: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*.

Almuhareb, Abdulrahman, and Poesio, Massimo. 2005. Concept Learning and Categorization from the Web. Pages 103–108 of: *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Altmann, Gerry T.M., and Kamide, Yuki. 1999. Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, **73**(3), 247–64.

Altmann, Gerry T.M., and Kamide, Yuki. 2007. The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, **57**(4), 502–518.

Andrews, Mark, Vigliocco, Gabriella, and Vinson, David P. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, **116**, 463–498.

Ashcraft, Mark H. 1978. Property norms for typical and atypical items from 17 categories: A description and discussion. *Memory & Cognition*, **6**, 227–232.

Aston, Guy, and Burnard, Lou. 1998. *The BNC handbook*. Edinburgh, UK: Edinburgh University Press.

Baker, Collin F., Fillmore, Charles J., and Lowe, John B. 1998. The Berkeley FrameNet Project. Pages 86–90 of: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*.

Balasubramanian, Niranjan, Soderland, Stephen, Mausam, and Etzioni, Oren. 2013. Generating Coherent Event Schemas at Scale. Pages 1721–1731 of: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*.

Barbu, Eduard. 2008. Combining methods to learn feature-norm-like concept descriptions. Pages 9–16 of: *Bridging the Gap between Semantic Theory and Computational Simulations: Proceedings of the ESSLLI 2008 Workshop on Distributional Semantics*.

Barbu, Eduard, and Poesio, Massimo. 2008. A Comparison of WordNet and Feature Norms. Pages 56–73 of: *Proceedings of the 4th Global Wordnet Conference (GWC 2008)*.

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

Baroni, Marco, Evert, Stefan, and Lenci, Alessandro (eds). 2008. *Bridging the Gap between Semantic Theory and Computational Simulations: Proceedings of the ESSLLI 2008 Workshop on Distributional Semantics*.

Baroni, Marco, Murphy, Brian, Barbu, Eduard, and Poesio, Massimo. 2010. Strudel: A Corpus-Based Semantic Model Based on Properties and Types. *Cognitive Science*, **34**, 222–254.

Bicknell, Klinton, Elman, Jeffrey L., Hare, Mary, McRae, Ken, and Kutas, Marta. 2010. Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, **63**(4), 489–505.

Bullinaria, John A., and Levy, Joseph P. 2007. Extracting semantic representations from word co-occurrence statistics: a computational study. *Behavior Research Methods*, **39**(3), 510–526.

Bullinaria, John A., and Levy, Joseph P. 2012. Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD. *Behavior Research Methods*, **44**(3), 890–907.

Chambers, Nathanael. 2013. Event Schema Induction with a Probabilistic Entity-Driven Model. Pages 1797–1807 of: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*.

Chambers, Nathanael, and Jurafsky, Daniel. 2008. Unsupervised Learning of Narrative Event Chains. Pages 789–797 of: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*.

Chambers, Nathanael, and Jurafsky, Daniel. 2009. Unsupervised Learning of Narrative Schemas and Their Participants. Pages 602–610 of: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*.

Cheung, Jackie Chi Kit, Poon, Hoifung, and Vanderwende, Lucy. 2013. Probabilistic frame induction. Pages 837–846 of: *Proceedings of the 2013 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Church, Kenneth Ward, and Hanks, Patrick. 1991. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, **16**(1), 22–29.

Collins, Allan M., and Loftus, Elizabeth F. 1975. A spreading-activation theory of semantic processing. *Psychological Review*, **82**, 407–428.

Collobert, Ronan, Weston, Jason, Bottou, Léon, Karlen, Michael, Kuksa, Pavel, and Kavukcuoglu, Koray. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, **12**, 2493–2537.

Cree, George S., McRae, Ken, and McNorgan, Chris. 1999. An attractor model of lexical conceptual processing: simulating semantic priming. *Cognitive Science*, **23**(3), 371–414.

De Deyne, Simon, Verheyen, Steven, Ameel, Eef, Vanpaemel, Wolf, Dry, Matthew J., Voorspoels, Wouter, and Storms, Gert. 2008. Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, **40**(4), 1030–48.

De Deyne, Simon, Verheyen, Steven, and Storms, Gert. 2015. The role of corpus size and syntax in deriving lexico-semantic representations for a wide range of concepts. *The Quarterly Journal of Experimental Psychology*, **68**(8), 1643–1664.

Devereux, Barry J., Pilkington, Nicholas, Poibeau, Thierry, and Korhonen, Anna. 2009. Towards Unrestricted, Large-Scale Acquisition of Feature-Based Conceptual Representations from Corpus Data. *Research on Language and Computation*, **7**(2-4), 137–170.

Devereux, Barry J., Tyler, Lorraine K., Geertzen, Jeroen, and Randall, Billi. 2014. The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior Research Methods*, **46**(5), 1119–1127.

Dowty, David. 1989. On the Semantic Content of the Notion of Thematic Role. Pages 69–130 of: Chierchia, Gennaro, Partee, Barbara H., and Turner, Raymond (eds), *Properties Types and Meaning*, Vol. II, Semantic Issues. Dordrecht; Kluwer Academic Publishers.

Dowty, David. 1991. Thematic Proto-Roles and Argument Selection. *Language*, **67**(3), 547–619.

Erk, Katrin. 2007. A Simple, Similarity-based Model for Selectional Preferences. Pages 216–223 of: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

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

Evert, Stefan. 2009. Corpora and Collocations. Chap. 58, pages 1212–1248 of: Lüdeling, Anke, and Kytö, Merja (eds), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.

Fagarasan, Luana, Vecchi, Eva Maria, and Clark, Stephen. 2015. From distributional semantics to feature norms: Grounding semantic models in human perceptual data. Pages 52–57 of: *Proceedings of the 11th International Conference on Computational Semantics*.

Fellbaum, Christiane. 1998. *WordNet - An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Ferretti, Todd R., McRae, Ken, and Hatherell, Andrea. 2001. Integrating Verbs, Situation Schemas, and Thematic Role Concepts. *Journal of Memory and Language*, **44**(4), 516–547.

Ferretti, Todd R., Kutas, Marta, and McRae, Ken. 2007. Verb Aspect and the Activation of Event Knowledge. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, **33**(1), 182–196.

Fillmore, Charles J. 1968. The Case for Case. Pages 0–88 of: Bach, Emmon, and Harms, Robert T. (eds), *Universals in Linguistic Theory*. New York: Holt, Rinehart and Winston.

Fort, Karën, Adda, Gilles, and Cohen, K. Bretonnel. 2011. Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*, **37**(2), 413–420.

Frermann, Lea, Titov, Ivan, and Pinkal, Manfred. 2014. A Hierarchical Bayesian Model for Unsupervised Induction of Script Knowledge. Pages 49–57 of: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*.

Garrard, Peter, Ralph, Matthew. A. Lambon, Hodges, John R., and Patterson, Karalyn. 2001. Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, **18**(2), 125–174.

Gildea, Daniel, and Jurafsky, Daniel. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, **28**(3), 245–288.

Greenberg, Clayton, Sayeed, Asad B., and Demberg, Vera. 2015. Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering. Pages 21–31 of: *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL.*

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery*. Dordrecht: Kluwer Academic Publishers.

Gruber, Jeffrey S. 1965. Studies in lexical relations. PhD thesis, Massachusetts Institute of Technology.

Hampton, James A. 1979. Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, **18**(4), 441 – 461.

Hare, Mary, Elman, Jeffrey L., Tabaczynski, Tracy, and McRae, Ken. 2009. The Wind Chilled the Spectators, but the Wine Just Chilled: Sense, Structure, and Sentence Comprehension. *Cognitive Science*, **33**(4), 610–628.

Harris, Zellig S. 1954. Distributional structure. *Word*, **10**, 146–162.

Hermann, Karl Moritz, Das, Dipanjan, Weston, Jason, and Ganchev, Kuzman. 2014. Semantic Frame Identification with Distributed Word Representations. Pages 1448–1458 of: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Hinton, Geoffrey E., and Shallice, Tim. 1991. Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, **98**, 74–95.

Jackendoff, Ray. 1972. *Semantic Interpretation in Generative Grammar*. Cambridge, MA: MIT Press.

Jackendoff, Ray. 1987. The Status of Thematic Relations in Linguistic Theory. *Linguistic Inquiry*, **18**(3), 369–411.

Kamide, Yuki, Altmann, Gerry T.M., and Haywood, Sarah L. 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, **49**(1), 133 – 156.

Kelly, Colin, Devereux, Barry, and Korhonen, Anna. 2010. Acquiring Human-like Feature-based Conceptual Representations from Corpora. Pages 61–69 of: *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*.

Kelly, Colin, Devereux, Barry J., and Korhonen, Anna. 2013. Automatic Extraction of Property Norm-Like Data From Large Text Corpora. *Cognitive Science*, **38**(4), 638–682.

Keuleers, Emmanuel, and Balota, David A. 2015. Megastudies, Crowdsourcing, and Large Datasets in Psycholinguistics: An Overview Of Recent Developments. *The Quarterly Journal of Experimental Psychology*, **68**(8), 1457–1468.

Kingsbury, Paul, and Palmer, Martha. 2003. PropBank: the Next Level of TreeBank. In: *Proceedings of Treebanks and Lexical Theories*.

Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation*, **42**(1), 21–40.

Kipper-Schuler, K. 2005. Verbnet: A broad-coverage, comprehensive verb lexicon. PhD thesis, University of Pennsylvania.

Kittur, Aniket, Nickerson, Jeffrey V., Bernstein, Michael, Gerber, Elizabeth, Shaw, Aaron, Zimmerman, John, Lease, Matt, and Horton, John. 2013. The Future of Crowd Work. In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*.

Kremer, Gerhard, and Baroni, Marco. 2011. A Set of Semantic Norms for German and Italian. *Behavior Research Methods*, **43**(1), 97–109.

Lapesa, Gabriella, and Evert, Stefan. 2014. A Large Scale Evaluation of Distributional Semantic Models: Parameters, Interactions and Model Selection. *Transactions of the Association for Computational Linguistics*, **2**, 531–545.

Lebani, Gianluca E. 2012. STaRS.sys: designing and building a commonsense-knowledge enriched wordnet for therapeutic purposes. PhD thesis, University of Trento.

Lebani, Gianluca E., Bondielli, Alessandro, and Lenci, Alessandro. 2015. You Are What You Do. An Empirical Characterization of the Semantic Content of the Thematic Roles for a Group of Italian Verbs. *Journal of Cognitive Science*, **16**(4), 401–430.

Lenci, Alessandro. 2008. Distributional semantics in linguistic and cognitive research. A foreword. *Italian Journal of Linguistics*, **20**(1), 1–30.

Lenci, Alessandro. 2011. Composing and Updating Verb Argument Expectations: A Distributional Semantic Model. Pages 58–66 of: *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*.

Lenci, Alessandro, Baroni, Marco, Cazzolli, Giulia, and Marotta, Giovanna. 2013. BLIND: A set of semantic feature norms from the congenitally blind. *Behavior Research Methods*, **45**(4), 1218–1233.

Levin, Beth, and Rappaport Hovav, Malka. 2005. *Argument Realization*. Cambridge, UK: Cambridge University Press.

Lin, Dekang. 1998. Automatic Retrieval and Clustering of Similar Words. Pages 768–774 of: *Proceedings of the 36th Annual Meeting of the Association for*

Computational Linguistics and 17th International Conference on Computational Linguistics.

Liu, Ding, and Gildea, Daniel. 2010. Semantic Role Features for Machine Translation. Pages 716–724 of: *Proceedings of the 23rd International Conference on Computational Linguistics*.

Manning, Christopher D., Raghavan, Prabhakar, and Schütze, Hinrich. 2008. *An Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press.

Màrquez, Lluís, Carreras, Xavier, Litkowski, Kenneth C., Stevenson, Suzanne (eds). 2008. *Computational Linguistics: Special Issue on Semantic Role Labeling*, **34**(2).

Matsuki, Kazunaga, Chow, Tracy, Hare, Mary, Elman, Jeffrey L., Scheepers, Christoph, and McRae, Ken. 2011. Event-Based Plausibility Immediately Influences On-Line Language Comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **37**(4), 913–934.

McRae, Ken, and Cree, George S. 2001. Factors underlying category-specific semantic deficits. In: Forde, E. M. E., and Humphreys, G. (eds), *Category Specificity in Mind and Brain*. Hove, East Sussex, UK: Psychology Press.

McRae, Ken, and Matsuki, Kazunaga. 2009. People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass*, **3**(6), 1417–1429.

McRae, Ken, De Sa, Virginia R., and Seidenberg, Mark S. 1997a. On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, **126**(2), 99–130.

McRae, Ken, Ferretti, Todd R., and Amyote, Liane. 1997b. Thematic Roles as Verb-specific Concepts. *Language and Cognitive Processes*, **12**(2/3), 137–176.

McRae, Ken, Spivey-Knowlton, Michael J., and Tanenhaus, Michael K. 1998. Modeling the Influence of Thematic Fit (and Other Constraints) in On-line Sentence Comprehension. *Journal of Memory and Language*, **38**(3), 283–312.

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

McRae, Ken, Cree, George S., Seidenberg, Mark S., and McNorgan, Chris. 2005b. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, **37**(4), 547–59.

Miller, George A., and Charles, Walter G. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, **6**(1), 1–28.

Montefinese, Maria, Ambrosini, Ettore, Fairfield, Beth, and Mammarella, Nicola. 2013. Semantic memory: A feature-based analysis and new norms for Italian. *Behavior Research Methods*, **45**(2), 440–461.

Nivre, Joakim, Hall, Johan, Nilsson, Jens, Chanev, Atanas, Eryigit, Gülsen, Kübler, Sandrs, Marinov, Svetoslav, and Marsi, Erwin. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, **13**(6), 95–135.

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

Padó, Ulrike. 2007. The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing. PhD thesis, Saarland University.

Palmer, Martha, Gildea, Daniel, and Xue, Nianwen. 2010. Semantic Role Labeling. *Synthesis Lectures on Human Language Technologies*, **3**(1), 1–103.

Poesio, Massimo, Barbu, Eduard, Giuliano, Claudio, and Romano, Lorenza. 2008. Supervised relation extraction for ontology learning from text based on a cognitively plausible model of relations. In: *Proceedings of the 3rd Workshop on Ontology Learning and Population*.

Pradhan, Sameer, Moschitti, Alessandro, Xue, Nianwen, Uryupina, Olga, and Zhang, Yuchen. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. Pages 1–40 of: *Joint Conference on EMNLP and CoNLL - Shared Task*.

Preiss, Judita, Briscoe, Ted, and Korhonen, Anna. 2007. A System for Large-Scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from Corpora. Pages 912–919 of: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*.

Recasens, Marta, Màrquez, Lluís, Sapena, Emili, Martí, M Antònia, Taulé, Mariona, Hoste, Véronique, Poesio, Massimo, and Versley, Yannick. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. Pages 1–8 of: *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*.

Regneri, Michaela, Koller, Alexander, and Pinkal, Manfred. 2010. Learning Script Knowledge with Web Experiments. Pages 979–988 of: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Reisinger, Drew, Rudinger, Rachel, Ferraro, Francis, Harman, Craig, Rawlins, Kyle, and Van Durme, Benjamin. 2015. Semantic Proto-Roles. *Transactions of the Association for Computational Linguistics*, **3**, 475–488.

Roller, Stephen, and Schulte im Walde, Sabine. 2014. Feature Norms of German Noun Compounds. Pages 104–108 of: *Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014)*.

Rosch, Eleanor, and Mervis, Carolyn B. 1975. Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology*, **7**, 573–605.

Roth, Michael, and Frank, Anette. 2013. Automatically Identifying Implicit Arguments to Improve Argument Linking and Coherence Modeling. Pages 306–316 of: *Second Joint Conference on Lexical and Computational Semantics (*SEM)*.

Roth, Michael, and Lapata, Mirella. 2015. Context-aware Frame-Semantic Role Labeling. *Transactions of the Association for Computational Linguistics*, **3**, 449–460.

Ruppenhofer, Josef, Gorinski, Philip, and Sporleder, Caroline. 2011. In Search of Missing Arguments: A Linguistic Approach. Pages 331–338 of: *Proceedings of Recent Advances in Natural Language Processing*.

Sahlgren, Magnus. 2006. The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in highdimensional vector spaces. PhD thesis, Stockholm University.

Sahlgren, Magnus. 2008. The distributional hypothesis. *Italian Journal of Linguistics*, **20**(1), 33–53.

Sartori, Giuseppe, and Lombardi, Luigi. 2004. Semantic relevance and semantic disorders. *Journal of Cognitive Neuroscience*, **16**(3), 439–52.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. Pages 44–49 of: *Proceedings of the international conference on new methods in language processing*.

Schulte im Walde, Sabine, and Melinger, Alissa. 2008. An in-depth look into the co-occurrence distribution of semantic associates. *Italian Journal of Linguistics*, **20**(1), 87–123.

Shen, Dan, and Lapata, Mirella. 2007. Using Semantic Roles to Improve Question Answering. Pages 12–21 of: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Smith, Edward E., Shoben, Edward J., and Rips, Lance J. 1974. Structure and Process in Semantic Memory: A Featural Model for Semantic Decisions. *Psychological Review*, **81**(3), 18–47.

Snow, Rion, O'Connor, Brendan, Jurafsky, Daniel, and Ng, Andrew Y. 2008. Cheap and Fast-but is It Good? Evaluating Non-expert Annotations for Natural Language Tasks. Pages 254–263 of: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Steyvers, Mark, Smyth, Padhraic, and Chemuduganta, Chaitanya. 2011. Combining Background Knowledge and Learned Topics. *Topics in Cognitive Science*, **3**(1), 18–47.

Storms, Gert, Navarro, Daniel J., and Lee, Michael D. (eds). 2010. *Acta Psychologica Special Issue on Formal Modeling of Semantic Concepts*. Vol. 133 (3).

Taylor, Wilson. 1953. Cloze Procedure: A New Tool for Measuring Readability. *Journalism Quarterly*, **30**, 415–433.

Tesnière, Lucien. 1959. *Eléments de Syntaxe Structurale*. Paris: Klincksieck.

Traxler, Matthew J., Foss, Donald J., Seely, Rachel E., Kaup, Barbara, and Morris, Robin K. 2001. Priming in Sentence Processing: Intralexical Spreading Activation, Schemas, and Situation Models. *Journal of Psycholinguistic Research*, **29**(6), 581–595.

Turney, Peter D., and Pantel, Patrick. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, **37**, 141–188.

Van Valin, Robert D. Jr. 1999. Generalized semantic roles and the syntax-semantics interface. Pages 373–389 of: Corblin, F., Dobrovie-Sorin, C., and Marandin, J.-M. (eds), *Empirical issues in formal syntax and semantics*. The Hague: Thesus.

Versley, Yannick, Ponzetto, Simone Paolo, Poesio, Massimo, Eidelman, Vladimir, Jern, Alan, Smith, Jason, Yang, Xiaofeng, and Moschitti, Alessandro. 2008. BART: A Modular Toolkit for Coreference Resolution. Pages 9–12 of: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*.

Vigliocco, Gabriella, Vinson, David P., Lewis, William D., and Garrett, Merrill F. 2004. Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, **48**(4), 422–88.

Vigliocco, Gabriella, Warren, Jane, Siri, Simona, Arciuli, Joanne, Scott, Sophie, and Wise, Richard. 2006. The role of semantics and grammatical class in the neural representation of words. *Cerebral Cortex*, **16**(12), 1790–1796.

Vinson, David P., and Vigliocco, Gabriella. 2008. Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, **40**, 183–190.

Vinson, David P., Vigliocco, Gabriella, Cappa, Stefano F., and Siri, Simona. 2003. The breakdown of semantic knowledge: Insights from a statistical model of meaning representation. *Brain and Language*, **86**, 347–365.

Wu, Shumin, and Palmer, Martha. 2011. Semantic Mapping Using Automatic Word Alignment and Semantic Role Labeling. Pages 21–30 of: *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*.

Zapirain, Benat, Agirre, Eneko, Màrquez, Lluís, and Surdeanu, Mihai. 2013. Selectional preferences for semantic role classification. *Computational Linguistics*, **39**(3), 631–663.