overall intent of the program. The language in which the programmer writes allows the sequence of events to be specified in much more general terms.

The models of language envisaged by Chomsky and his followers are similar to computers in requiring steps to be specified in minute detail. They differ to the extent that computer scientists recognize that there is a point beyond which detail obscures more than it reveals. Mathematicians do not attempt to explain the notion of quotients by describing the process of long division. Indeed, the only way a student can reasonably hope to understand long division is to first get a firm grasp of quotients. The possibility of bypassing long division that computers and calculators provided was seized immediately, and now we hear of it no more.

Computational linguistics has been, in large measure, an attempt to bring to linguistics some of the advantages that computers allowed in the teaching of arithmetic and that compilers brought to computer science. Given that the field is barely 60 years old, its successes have been remarkable. It has been successful in providing a number of models of linguistic processing whose computational power matches what the task seems to require extremely closely. It is powerful enough to allow all the necessary computation to be specified but constrained enough to capture important insights into the fundamental nature of linguistic processes.

Computational linguistics has given us a version of finite-state technology that is apparently a remarkably good fit for the requirements of morphology and morpho-phonology. It does not require the linguist who uses it to think in terms of states and transitions, but rather in terms of a sequence of representations of a word, each related in a straightforward way to the ones immediately preceding and following in the sequence, connecting an abstract representation at one end of the sequence with the observed representation at the other. This has become a common way of thinking among formal linguists, who generally find it very congenial.

More striking are the contributions that computational linguists have made in syntax, where such formalisms as Lexical Functional Grammar, Head Driven Phrase Structure Grammar, and Combinatory Categorial Grammar are widely acknowledged as major contributions. In these frameworks, Chomsky's problem of how to form a question from the sentence *The man who is hungry is ordering dinner* simply does not arise. The assertion and the question are indeed assumed to have similar underlying structures, but neither has a privileged position relative to those structures.

5.4 Conclusion

Nontrivial tasks, like translation, are AI-complete. In other words, any machine that can perform them would have to be able to mimic all aspects of human

intelligence. This is nowhere more evident than in machine translation. At no time since the launch of Sputnik caused serious work to begin on this problem has it been thought important that workers in this field should inform themselves about what a translation is generally taken to be by those who use them or what it is that distinguishes a professional translator from someone who took some French at school. It was, however, thought to be useful to know something of what linguists have discovered about language. Not only is any such knowledge now thought to be unnecessary; it is often regarded as a hindrance. With such knowledge, we will reject the possibility that *He had eaten earlier that evening* might, on rare occasions, be rendered into Spanish as *No había comido esa misma tarde*, and certainly that *man bites dog* is a possible, if rare, equivalent of *dog bites man*.

As we have already noted, one of the most remarkable properties of human language is that, despite its great subtlety and complexity, it is a tool that humans can use in an entirely casual manner. Communication in this medium is a collaborative enterprise in which the receiver is constantly second-guessing the sender's intentions. It is, therefore, a fundamentally probabilistic enterprise. When we say that nontrivial linguistic tasks are AI-complete, we are also saying that they are probabilistic. During most of the time since the launch of Sputnik, probability and statistics have been largely absent from linguistics. This is unfortunate and it doubtless did much to hamper progress in the field. But, slowly and cautiously, it is a matter that is being set to rights. There are now rooms in university linguistics departments with signs on the door saying things like *Syntax Laboratory*. Linguists use computer tools to search large corpora for examples, and the judgments of individual informants are no longer taken as sacred truth. But linguists have not embraced big data unquestioningly or replaced thought and analysis with machine learning. We must hope they never do.

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

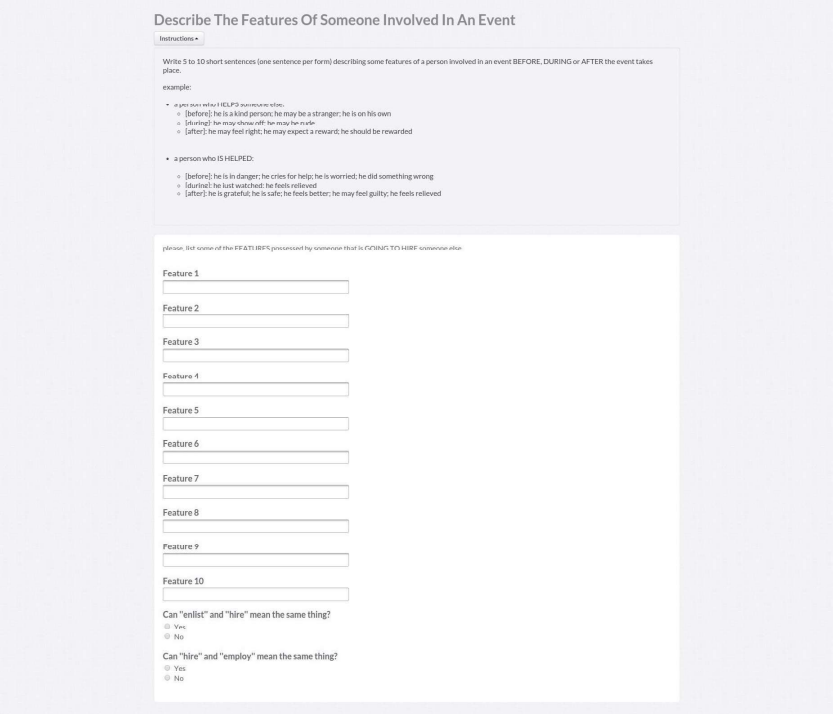

Figure 6.1 The verb role description interface in Crowdflower.

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:

$$\mathrm{L} = \langle admit, subject\rangle, \langle plead, subject\rangle, \langle convict, object\rangle, \langle sentence, object\rangle$$

$$\mathrm{O} = \{(plead, convict), (convict, sentence), ...\}$$

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

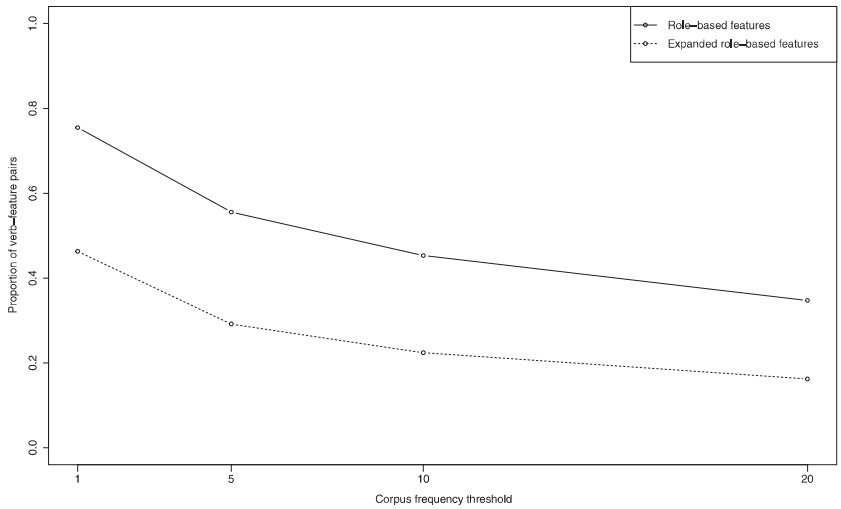

Figure 6.2 Proportion of verb-feature pairs in a ±2-sentences window, modulated by application of different frequency thresholds.

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 ±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

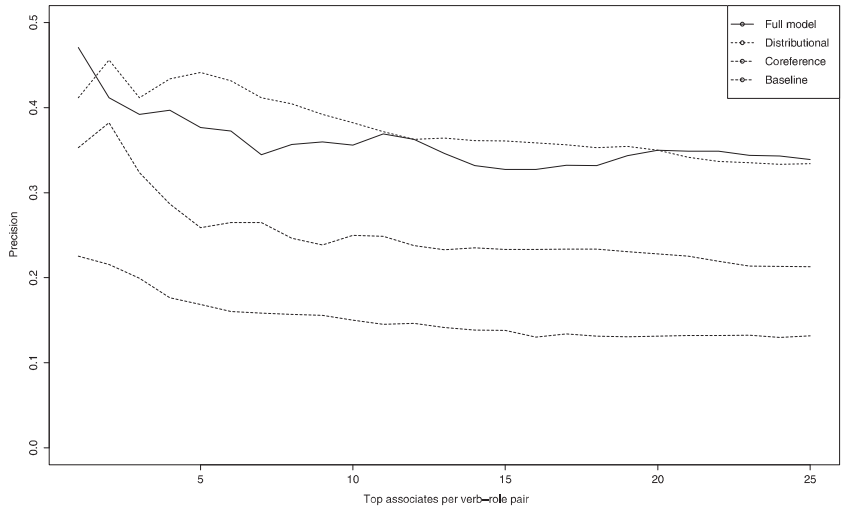

Figure 6.3 Precision of the different models evaluated against the dataset of extended role-based features.

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,

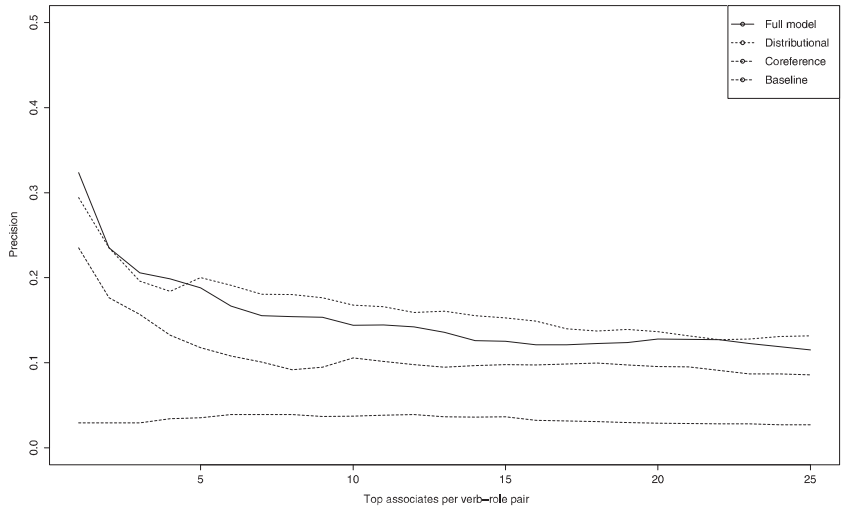

Figure 6.4 Precision of the different models evaluated against the dataset of role-based features.

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.

7 Native Language Identification on EFCAMDAT

Xiao Jiang, Yan Huang, Yufan Guo, Jeroen Geertzen, Theodora Alexopoulou, Lin Sun, and Anna Korhonen

Abstract

Native Language Identification (NLI) is a task aimed at determining the native language (L1) of learners of second language (L2) on the basis of their written texts. To date, research on NLI has focused on relatively small corpora. We apply NLI to EFCAMDAT, an L2 English learner corpus that is not only multiple times larger than previous L2 corpora but also provides pseudo-longitudinal data across several proficiency levels. Based on accurate machine learning with a wide range of linguistic features, our investigation reveals interesting patterns in the longitudinal data that are useful for both further development of NLI and its application to research on L2 acquisition.

7.1 Introduction

Native language identification (NLI) is a task aimed at detecting the native language (L1) of writers on the basis of their second language (L2) production. NLI is important for natural language processing (NLP) applications including language tutoring systems and authorship profiling. Moreover, NLI can offer useful empirical data for research on L2 acquisition. For example, NLI can shed light on how L1 background influences L2 learning, and on differences between the writings of L2 learners across different L1 backgrounds.

To date, studies on NLI have focused on relatively small learner corpora. Furthermore, none of them have investigated the influence of L1s across L2 proficiency levels. Our work takes the first step toward addressing these problems. We apply NLI to EFCAMDAT, the EF-Cambridge Open Language Database (Geertzen, Alexopoulou, and Korhonen, 2013),[1] an open-access L2 learner corpus.

EFCAMDAT consists of writings of learners submitted to *Englishtown*, the online school of EF. EFCAMDAT stands out for its size, diversity of student

[1] http://corpus.mml.cam.ac.uk/efcamdat

backgrounds, and coverage of the proficiency levels. The first release of 2013 (Geertzen, Alexopoulou, and Korhonen, 2013), on which this paper is based, amounts to 30 million words, a corpus multiple times larger than any other available L2 corpora. Using a standard machine learning–based methodology for NLI, we explore the optimal linguistic features for NLI on this data at different proficiency levels. We discover interesting patterns that can be useful for both further development of NLI and its application to research on L2 acquisition.

In this introductory section, we first review the history of research on NLI, and introduce the data sets that have been used in earlier NLI research. We then summarise our contribution briefly. Section 7.2 describes our data set EFCAMDAT in detail, Section 7.3 describes our research method, Section 7.4 presents our empirical results and qualitative analysis, and finally Section 7.5 presents our conclusions.

7.1.1 Prior studies on NLI

The first study on NLI was conducted by Tomokiyo and Jones (2001). While their original goal was to develop techniques for detecting nonnative speech, they actually built a Naive Bayes classifier to distinguish between the native speakers of Chinese and Japanese according to the transcripts of their English utterances. Based on word n-grams in which nouns were replaced by their part-of-speech (POS) tags, the classifier achieved a remarkable accuracy of 100%. However, the generalizability of the result is questionable considering the limited population of the subjects: only eight English and six Chinese speakers were involved.

Koppel, Schler, and Zigdon (2005) regarded NLI as a subtask of authorship attribution, and constructed a Support Vector Machine (SVM) classifier for NLI on five native language backgrounds. With 1,035 features, including function words, character n-grams, error types extracted by the grammar checker of Microsoft Word, rare words, and POS n-grams, the classifier achieved an accuracy of 80.2%. The study became a benchmark for most subsequent research.

Tsur and Rappoport (2007) replicated the study of Koppel, Schler, and Zigdon (2005), but sampled the texts differently and fed only one type of features to the classifier at a time. They found that character bi-grams were most discriminatory for NLI; the highest classification accuracy was 66%. The accuracy of bi-gram classifier remained high even when the dominant content words or function words were removed from the texts. It was concluded that the word choice of second-language writing might be subject to the phonology of writers' native languages.

Wong and Dras (2009) also followed the study of Koppel, Schler, and Zigdon (2005). They extended the number of native language backgrounds in

NLI to seven. In addition to the feature sets of Koppel, Schler, and Zigdon (2005), they studied the discriminatory power of three types of syntactic errors extracted by an automatic grammatical checker. However, results showed that the effect of these errors was not prominent, which might be attributed to the high false-positive rate of the grammatical checker, as well as the limited types of the errors. The highest classification accuracy with combined features was 73.71%. Nevertheless, the value of learner errors for NLI was highlighted by Kochmar (2011) later. He demonstrated that the error types which were typical in the English writings by native speakers of various Indo-European languages were useful for pairwise classification of these native language backgrounds. The accuracy of error-based classifiers ranges from 47.92% to 59.98%.

In a subsequent study, Wong and Dras (2011) introduced two types of syntactic features to NLI: one consists of production rules extracted from Context Free Grammar (CFG) parse tree, while the other contains reranking features previously used to select the best tree derivatives returned by parsers (Charniak and Johnson, 2005). These features led to a remarkable reduction of 30% in classification errors over the baseline. The highest classification accuracy was 81.71%. Based on the same data, Wong, Dras, and Johnson (2011; 2012) continued to investigate novel features, which included topical features extracted by Latent Dirichlet Allocation (LDA) and mixed POS and function word n-grams sifted by adaptor grammar; the latter were found to be useful in NLI. Combined with mixed n-grams containing POS and function words, it achieved a classification accuracy of 75.71%.

A number of studies were then conducted following the same data setting of Wong and Dras (2011). Ahn (2011) found that the effect of character tri-grams was highly correlated with that of word uni-grams, and that their contribution to the NLI classification accuracy was attributable to topic biases. Bykh and Meurers (2012) studied how the effect of word and POS n-grams on NLI changed with their length. They accomplished a classification accuracy of 89.71% using recurring n-grams, which were n-grams that appeared in more than two texts in sets. Swanson and Charniak (2012) explored the effect of a novel feature type, Tree Substitution Grammar (TSG) fragments, and found that the classifiers built on such features outperformed those built on CFG production rules (78.4% versus 72.6% in accuracy).

Jarvis and Crossley (2012) conducted a series of research into NLI aiming at detecting L1 transfer. With Latent Discrimant Analysis (LDA), they systematically investigated both the independent and combined effects of word n-grams, Coh-Metrix stylistic features (Graesser et al., 2004), and manually annotated errors in NLI. Their results also justified the effect of errors: when different feature types were fed independently to the classifier, the fine-grained manually annotated errors led to the highest three-way classification accuracy of 65.5%.

Tetreault et al. (2012) trained classifiers for separate feature types, and combined their outputs to build an ensemble classifier that produced the highest score so far for seven-way NLI (90.1%) on the commonly used International Corpus of Learner English (ICLE) (Granger, 2003). The ICLE data used by Tetreault et al. (2012) underwent some corrective processing, and were controlled for topic distribution, so they were not entirely the same as those used in earlier studies. They also conducted cross-corpus evaluation, and found that the classifiers trained on the small ICLE data cannot generalize well to larger corpora, while those trained on the larger corpora can generalize to the ICLE data.

As a response to an increased research interest in NLI, the first NLI shared task was held in 2013 (Tetreault, Blanchard, and Cahill, 2013), attracting 29 participating teams. A number of novel features were tried, such as the round-trip translation of English words which were intended to capture lexical preferences of each native language group (Lavergne et al., 2013), brown clusters that were produced by hierarchical clustering of words based on the context, and restored tags that were achieved by removing some words from text and recovering the words according to an n-gram language model to capture consistent omission or misuse of these words (Tsvetkov et al., 2013). However, most of these new features brought marginal or no empirical improvement to NLI. The most successful teams (Jarvis, Bestgen, and Pepper, 2013) generally involved long n-grams of words (at least quad-grams) and characters (up to 9-grams) in building their classifiers. The system that ranked third in the closed NLI shared task was solely built on characters using Kernel Ridge Regression (KRR) (Popescu and Ionescu, 2013). By adding an intersection kernel in a follow-up study, the same team achieved an accuracy that is 1.7% higher than the top scoring team in the 2013 shared task (Ionescu, Popescu, and Cahill, 2014).

While the aforementioned studies only focused on English L2 data, Malmasi and Dras extended the study of NLI to Arabic L2 (Malmasi and Dras, 2014a) and Chinese L2 (Malmasi and Dras, 2014b). They achieved accuracies of 41.0% and 70.61% on these languages, respectively, using function words, context-free grammar production rules, and POS n-grams as features. Furthermore, the authors found the classification accuracies on Chinese and English L2 data were similar.

There has also been research into comparing different classifiers for NLI. The results have been mixed. During an author-profiling study in which the native language was also one of the dependent variables, Estival et al. (2007) compared a number of classifiers including decision trees, SVM and ensemble learning, etc., and found that the Random Forest (RF) classifier with feature selection based on information gained best results for NLI on an e-mail data set. Jarvis (2011) studied 20 classifiers on NLI and found LDA to be the most

accurate, while RF and commonly used SVM classifiers were substantially worse. However, it should be noted that their study only used word n-grams as features, and that the best system in the 2013 shared task employed SVM (Jarvis, Bestgen, and Pepper, 2013).

7.1.2 Prior Data Sets

In terms of data sets, most previous studies on NLI employed the International Corpus of Learner English (ICLE) corpus (Granger, 2003), developed at the Centre for English Corpus Linguistics at the University of Louvain, Belgium.[2] The ICLE corpus was specifically designed for the study of English writings from nonnative speakers. It includes English writings from university students worldwide. These students were roughly at the higher-intermediate or advance English proficiency level. The corpus contains several subcorpora, each of which contains writings of students of one of the following native languages: Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, and Swedish. In total, there are 3,640 writings and 2.5 million words in the corpus. Most writings are argumentative essays or literature examination papers. Example topics are "Crime does not pay" and "The role of censorship in Western society." Each writings must be at least 500 words long. On average there are 690 words per writing.

While ICLE has been widely used in NLI studies (Koppel, Schler, and Zigdon, 2005; Tsur and Rappoport, 2007; Wong and Dras, 2009; Wong and Dras, 2011; Ahn, 2011; Bykh and Meurers, 2012; Jarvis and Crossley, 2012), it is subject to topic biases (Ahn, 2011; Brooke and Hirst, 2011; Tetreault et al., 2012). Since the topics for the essays in ICLE were chosen individually by each university involved in the project, some topics are relevant to students only from a specific native language background. For example, many of the common topics in the French subcorpus concern the relatively esoteric subjects of literature, religion, and politics, while most of the Japanese subcorpus consists of more personal topics, ranging from experience as an English learner to one's favorite travel destination. Thus, it is inevitable that the writings by French authors appear to be more formal while those by Japanese authors appear to be more narrative and colloquial. In this case, the classification accuracy of native language identification is conflated by that of topic classification. Brooke and Hirst (2011) demonstrated the topic biases in ICLE through cross-corpus evaluation: they trained a classifier on ICLE data, and applied it to a new corpus Lang-8[3] that contains 22-million-word short journal entries covering a range of topics; the classification accuracy dropped by more than 54%. Brooke and Hirst (2011) believed that the topic biases in ICLE were the cause of the sharp drop, though

[2] http://www.uclouvain.be/en-317607.html [3] http://lang-8.com

the drop may also have been aggravated by the heterogeneous and incoherent nature of Lang-8 (Bykh and Meurers, 2012). Meanwhile, there is other empirical evidence of the topic biases in ICLE: Ahn (2011) demonstrated that when excluding topic words unique to each native language subcorpus, the performance of classifier based on uni-grams deteriorated dramatically.

Also, the distribution of L2 proficiency in ICLE was found to be unbalanced across different native language groups, which could also conflate the classification accuracy of NLI. Bestgen, Granger, Thewissen, et al. (2012) rated their data from ICLE in the Common European Framework (CEF), and found that there was significant difference in the English proficiency level of the French, German, and Spanish groups, and that the occurrence of learner errors was negatively correlated with the L2 proficiency level of the writers. For research that is intended to analyze L1 transfer, the effect of L2 proficiency level has to be controlled.

Some prior NLI studies have also used other data sets. Al-Rfou (2012) exploited English Wikipedia comments in their research. As 47% of the 60,000 Wikipedia editors had specified their native languages as non-English, the editor comments could be used for NLI. Al-Rfou (2012) built a corpus that contained more than 12 million words from 9,857 distinct authors representing the top 20 most frequently used native languages. They argued that working on the Wikipedia data set was more challenging because each individual writing was shorter, and that the writings were diverse in topics as well as the English proficiency levels of their writers. Kochmar (2011) used Cambridge Learner Corpus, which consists of more than 200,000 writings produced during Cambridge ESOL English exams by students from 217 countries. As stated earlier, the corpus has manual error tags, which enabled Kochmar (2011) to study the usefulness of error-driven features in NLI. Estival et al. (2007) used 9,836 English-language e-mails written by English, Arabic, and Spanish people. Jarvis and Crossley (2012) utilized narrative film descriptions on only one topic for studying the effect of uni-grams. Tetreault et al. (2012) adopted TOEFL 11 corpus, which contained 11,000 essays written by test-takers during The Test of English as a Foreign Language (TOEFL), with 1,000 texts for each of 11 native language groups. The corpus was then enlarged to 1,100 texts for each group (Blanchard et al., 2013), and was used in the closed NLI shared task in 2013 (Tetreault, Blanchard, and Cahill, 2013).

7.1.3 Our contribution

In this study we employ the recently released EFCAMDAT corpus for NLI. Since the data come from a live educational context, where the writings are submitted by a large number of students from diverse backgrounds, the corpus provides a rich resource for the development and evaluation of NLI as well as for linguistic studies. We explore the potential and challenges of NLI when applied

to this new longitudinal data set. We report experiments where a rich set of linguistic features were extracted from this corpus (including word and character n-grams, POS n-grams, production rules, and grammatical relations) using state-of-the-art NLP and classified using Support Vector Machines (SVM). We conduct a quantitative and qualitative evaluation of the performance of different features and compare, for the first time, the performance of different features at different proficiency levels. We observe patterns interesting for the development of both NLI and L2 acquisition research: the top performing features differ across proficiency levels and such patterns can have relevance for research on L2 acquisition.

7.2 Data

EFCAMDAT was developed at the University of Cambridge, in collaboration with Education First (EF) – an international organization of teaching English as a foreign language. The corpus consists of writings submitted to EF Englishtown,[4] the online school of EF. The EF curriculum is organised along 16 teaching levels, which are aligned to the Common European Framework of Reference Levels (CEFR). EF teaching levels 1–3 correspond to CEFR A1, 4–6 to CEFR A2, 7–9 to CEFR B1, 10–12 to CEFR B2 and 13–16 to C1. Each teaching level consists of eight lessons; at the end of each lesson there is a writing assignment that learners submit for correction by (human) EF teachers. When learners sign up for a course, they take a placement test to identify their initial proficiency level. The majority of learners in EFCAMDAT complete one to three teaching levels. For further details, see Geertzen, Alexopoulou, and Korhonen (2013).

Each writing is accompanied with metadata including its submission date, EF teaching level, teaching unit and lesson title, topic/task ID, as well as a grade marked by a human grader from EF; metadata for authors include an anonymous id, country, and nationality. A combination of nationality and country have been used as a proxy to the native language background.

Table 7.1 shows the current total number of documents, words, learners, nationalities, and proficiency levels covered by EFCAMDAT (as of October 2013).

Table 7.2 shows the top 10 nationalities with most writings, from 175 different nationalities. All nationalities in the Top 10 list have their unique native languages. The native language of Brazilians is Portuguese, and Spanish is that of Mexicans. Writings from these 10 nationalities make up 90% of the whole corpus.

[4] http://www.englishtown.com/

Table 7.1 *The statistics of* EFCAMDAT

	Count
Documents	423,373
Words	30,763,521
Learners	76,002
Nationalities	175
Proficiency levels	16

Table 7.2 *Top 10 nationalities by the number of writings*

Rank	Nationality	# Writings	Rank	Nationality	# Writings
1	Chinese	162,256	6	Mexican	15,802
2	Brazilian	71,182	7	French	14,067
3	Russian	63,470	8	Saudi Arabians	7,743
4	Italian	19,304	9	Americans	6,712
5	German	17,030	10	Japanese	6,027

Table 7.3 shows the number of writings at each of the 16 different proficiency levels. As can be seen, most writings come from low proficiency levels. Writings from levels 1 to 3 make up 41% of the whole corpus; writings from levels 4 to 7 make up another 40% of the whole corpus; thus, almost 80% of writings fall broadly within CEFR level A.

There are 128 tasks following topics such as the ones shown in Table 7.4.

EFCAMDAT is different from the frequently used ICLE corpus in many ways. First, EFCAMDAT is an order-of-magnitude larger than ICLE. The former contains half a million writings from 76,000 authors that total 30 million words, while the latter only contains 3,640 writings totaling 2.5 million words. As new

Table 7.3 *Number of writings at 16 proficiency levels*

Lvl	# Writings	Lvl	# Writings
1	91,948	9	16,056
2	38,220	10	21,710
3	43,776	11	10,180
4	74,700	12	5,695
5	35,922	13	3,985
6	20,214	14	1,567
7	39,700	15	745
8	18,456	16	499

Table 7.4 *Example topics for 16 proficiency levels*

Lvl	Example Topics	Lvl	Example Topics
1	Giving instructions to play a game	9	Giving feedback to a restaurant
2	Writing a birthday invitation	10	Helping a friend find a job
3	Renovating your home	11	Writing a movie review
4	Describing your family in an e-mail	12	Entering a writing competition
5	Giving suggestions about clothing	13	Writing a campaign speech
6	Writing a movie plot	14	Applying for sponsorship
7	Writing a letter of complaint	15	Covering a news story
8	Making a dinner party menu	16	Criticizing a celebrity

data keep coming in, the size of EFCAMDAT continues to grow. Second, the authors in EFCAMDAT are more diverse. They come from 175 different countries and represent 16 different proficiency levels. In contrast, the authors of ICLE are college students at roughly the same age worldwide and have advanced English proficiency. The average writing length in ICLE (500–1,000 words per writing) is much longer than that in EFCAMDAT (50–120 words per writing). Third, the topics are diverse in EFCAMDAT. There are 128 different task prompts across the 16 proficiency levels involving a variety of narrative, descriptive, and argumentative tasks from everyday communication such as "Introducing yourself by e-mail" (Level 1) to more complex tasks like "writing a movie plot" (Level 6). In contrast, most writings in ICLE are argumentative essays or literature examination paper. Being a collaboration project with many international universities, ICLE granted each university the right to assign its own topics. Therefore, the topics vary across different L1 language groups. In EFCAMDAT, the topics set for the writings at certain proficiency level are the same across all L1 backgrounds.

Performing NLI on EFCAMDAT is challenging. Being a real-world data set from EF school, the huge diversity of learner backgrounds makes the corpus noisy. Also, the average length for each writing is much shorter, and we can extract less information per writing in classification experiments. Moreover, the writings of EFCAMDAT have a relatively limited vocabulary and structure. Each writing corresponds to a lesson, and the students tended to replicate the vocabulary and sentence structure they just learnt from that lesson. Therefore, many writings from the same topic may present similar vocabulary and structure, even though they come from different learners with different native language backgrounds. Such similarity makes it more difficult to distinguish writings among different native language backgrounds.

Since we wanted to investigate different proficiency levels and not all levels had sufficient data for adequate NLI performance at the time of this experiment, we merged proficiency levels into four groups as to avoid data sparsity, as shown in Table 7.5. In doing this we have largely respected the alignment of

Table 7.5 *A subset of* EFCAMDAT *used in this work*

Group	Documents	Words
Lvl 1–3	44,362	1,910,674
Lvl 4–7	50,593	4,223,560
Lvl 8–11	15,095	1,726,093
Lvl 12–16	2,459	359,563

EF teaching levels to CEFR: levels 1–3 correspond to CEFR A1; the grouping 4–7 broadly aligns with CEFR A2 even though it includes EF 7 which corresponds to early B1; levels 8–11 align with the intermediate CEFR level B1/B2; while 12–16 broadly align with CEFR advanced C. We focused on three major nationalities – Chinese, Brazilian, and Russian – which jointly cover 78% of the corpus and yield a reasonably large training and test sets for NLI. We excluded some essays to ensure that each group contains approximately the same amount of data.

7.3 Methods

This section describes the methodology for building our NLI system on EFCAMDAT.

7.3.1 Features

We investigated a variety of lexical and syntactic features used in previous NLI works:

Word n-gram The sequence of adjacent words with length n. For example, for sentence `I speak English`, the word bi-gram (n = 2) features are `I speak` and `speak English`.

We experimented with word n-grams of different orders (n = 1 to 4) under different settings. These settings included whether to convert all letters to lower case (LC), to perform stemming (STM), to remove stop words and punctuation (RMSTP), to filter out nonalphanumeric words (ALPHANUM), and to filter out n-grams that were less frequent (MF) in the whole data set, or n-grams that had low recurring frequencies across different writings (RMF), i.e., appearing in fewer than a threshold number of different writings. LC, STM, RMSTP, and ALPHANUM were supposed to prevent overly specific features that may lead to data sparsity issues. MF and RMF were supposed to discard any rare features that were less informative and were potential noises to classifiers. The function word

features mentioned in many previous literatures were treated as another special setting for word uni-gram – only function words were kept as features while any other n-grams were discarded.

Character n-gram The sequence of adjacent characters of length n. For example, for sentence `I speak English`, some of the character tri-gram (n = 3) features are `I s`, `spe`, `eak`, `k E`, etc.

The character n-grams can span across word boundaries and measure the sentence in different granularity by using different order n. Character n-grams of small n may capture phonotactics, morpheme (the smallest unit of a word) information, prefix and suffix usage (Tsur and Rappoport, 2007), and even some spelling errors, while those of large n may capture the whole word characteristics (Ahn, 2011). We experimented with character n-grams in various settings as in that for word n-grams.

POS n-gram The sequence of part-of-speech (POS) tags for adjacent words of length n.

The POS tags indicate the syntactic or morphological category of words. The Penn Treebank POS tag set (Mitchell Marcus, 2012) was used. For example, `VBP` stands for non–third-person singular present verb, `PRP` stands for personal pronoun, and `NNP` for singular proper noun. Therefore, the example POS bi-grams for the same sentence `I speak English` are `PRP VBP` and `VBP NNP`. We experimented POS n-grams of different orders (n = 2 to 5). Inspired by the study of Bykh and Meurers (2012), we adopted three different types of POS n-grams (Table 7.6).

The pure POS feature uses only POS tags without lexical information. The POS+Word feature is the lexicalized POS n-gram, binding the words to their POS tags. The hybrid feature replaces all open-class words, e.g., nouns, verbs, adjectives, etc., with their POS tags to eliminate content-dependent information. Each of the three types of POS n-grams was experimented with different orders (n = 2 to 5). For hybrid POS n-grams, we also replaced different word classes and compared the results. By doing that we could achieve a clear understanding about which word class contributed most to the classification performance.

Table 7.6 *Example of three POS n-gram subtypes*

Subtypes	Description	Example
Pure POS	POS tags only	[PRP, VBP, DT, NN]
POS + Word	Lexical words and their POS tags	[I_PRP, drink_VBP, the_DT, water_NN]
Hybrid	Replace open-class lexical words (i.e., nouns, verbs, adjectives) with their POS tags	[I VBP the NN]

Parse Tree Output:

```
(ROOT
 (S
   (NP (PRP I))
   (VP (VBP speak)
     (NP (NNP English)))))
```

Unlexicalized PR:	**Lexicalized PR:**
S -> NP + VP	S -> NP + VP
NP -> PRP	NP -> PRP_I
VP -> VBP + NP	VP -> VBP_speak + NP
NP -> NNP	NP -> NNP_English

Figure 7.1 Example of unlexicalized and lexicalized version of production rule features (PR) of sentence "I speak English".

Production rule The rewriting rule that specifies a symbol substitution for generating a new symbol sequence.

A production rule is extracted from a sentence parse tree under the framework of Context Free Grammar (CFG). For example, from the sentence `I speak English`, we can extract the production rule S→NP+VP, which stands for the structure of a noun phrase NP (`I`) and a verb phrase VP (`speak English`). In addition to standard production rules, we also tested lexicalized production rules, with the corresponding words attached to each symbol, e.g., *S→NP_I+VP_went*. Figure 7.1 demonstrates the parse tree for a sample sentence and the two versions of production rule features respectively.

Dependency The functional relationship between a pair of words, where one word is the head and the other is the dependent.

In the example sentence, `dobj(speak, English)` is a dependency relation, which denotes that `English` is the direct object (`dobj`) of the verb `speak`. Such feature can capture the relations between noncontiguous word pairs. We adopted Stanford typed dependency scheme (SD), and obtained the dependencies by converting constituency parse trees with heuristic rules (De Marneffe, MacCartney, and Manning, 2006).

We experimented with both lexicalized and unlexicalized dependencies. Figure 7.2 demonstrates the dependency parse of sentence *I come from western Europe* and the two versions of dependency features.

7.3.2 Machine Learning Techniques

We tried two popular classifiers in text classification: Naive Bayesian (NB) and Support Vector Machine (SVM) for NLI. The result showed that the latter

Dependencies Output:

```
nsubj(come-2, I-1)
root(ROOT-0, come-2)
prep(come-2, from-3)
amod(Europe-5, western-4)
pobj(from-3, Europe-5)
```

Unlexicalized Depd:	**Lexicalized Depd:**
nsubj	nsubj_come_I
prep	prep_come_from
amod	amod_Europe_western
pobj	pobj_from_Europe

Figure 7.2 Example of unlexicalized and lexicalized version of dependency features (Depd) on sentence "I come from western Europe".

significantly outperformed the former in terms of overall accuracy. In a preliminary seven-way classification experiment where the data set size was 40K and the dimension of feature vector was 14K, SVM achieved an accuracy of 61% whereas NB achieved an accuracy of only 49%. In view of this, in the rest of this chapter, we only report the result from SVM classifier.

Support vector machine (SVM), having yielded the best performance in many NLP studies, is one of the most popular machine learning techniques in the field. A SVM classifier finds a separating hyperplane in the high-dimensional vector spaces of a given data set. A good hyperplane should have the largest distance to any data point of any class. The largest distance is termed as margin.

Given a data set $S = [(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)]$, the hyperplane that maximizes the margin is obtained by finding the hyperplane defined by $<\mathbf{w}, b>$ that

minimizes: $\frac{1}{2} \|\mathbf{w}\|^2$
subject to constrain: $y_i(\mathbf{w}x_i - b) \geqslant 1$

In the case where samples are not linearly separable, a kernel trick is adopted to project the data points into higher dimensional space where the data points are linearly separable. The inner products between high-dimensional data points can then be computed efficiently.

Most prior studies on NLI used SVM for classification (Koppel, Schler, and Zigdon, 2005; Tsur and Rappoport, 2007; Wong and Dras, 2009; Wong and Dras, 2011; Kochmar, 2011; Al-Rfou, 2012; Bykh and Meurers, 2012; Tetreault, Blanchard, and Cahill, 2013). It can suit the unique properties of text classification: the feature vector dimension is usually many orders of magnitude higher than other classification tasks; the features are highly relevant across data set; sometimes they are linearly inseparable. These properties make

SVM suitable for text classification, which has also been attested by Joachims (1998).

SVM is originally designed for binary classification. However, for NLI, a multiclass classifier is needed. A number of strategies have been proposed for combining multiple binary SVMs to build a multiclass classifier (Bishop, 2006). The most common approach is to construct a set of SVMs for each individual class, where each SVM is trained using data from one class as positive examples and the rest of the classes as negative examples. Predications for new inputs are made by choosing the class corresponding to the greatest margin. This is known as the one-versus-the-rest approach, which usually gives good results (Vapnik, 1998), and is the strategy used by LIBLINEAR (Fan et al., 2008) – the machine learning library that we used in our experiments.

7.3.3 Experiment Setup

We used the Stanford parser (Klein and Manning, 2003) to extract syntactic features. Following Bykh and Meurers (2012), we used the LibLinear SVM classifier (Fan et al., 2008) for NLI. To avoid selection bias, we performed fourfold cross-validation and reported the average accuracy at levels 1–3, 4–7, 8–11, and 12–16, respectively.

7.4 Results

First, we investigated the performance of individual feature types. We then investigated the combined effect of the features. Furthermore, we extracted the most distinguishing features and qualitatively analyzed these features with respect to the L1 backgrounds of the L2 learners. In general, our results show that lexical features (word and character n-grams) achieve higher classification accuracy than grammatical features (POS n-grams, production rules and dependencies). Nevertheless, as the proficiency level grows, the performance of lexical features drops while the performance of grammatical ones grows. This may imply that as students progress in L2, they use more complicated grammatical structures that show influence from their native languages.

7.4.1 Individual Features

This section presents our findings of the best settings for individual feature types in NLI and compares the accuracy of NLI across different individual feature types.

Effect of different feature settings In this section, we explore the effect of different feature settings on classification accuracy. For word n-grams,

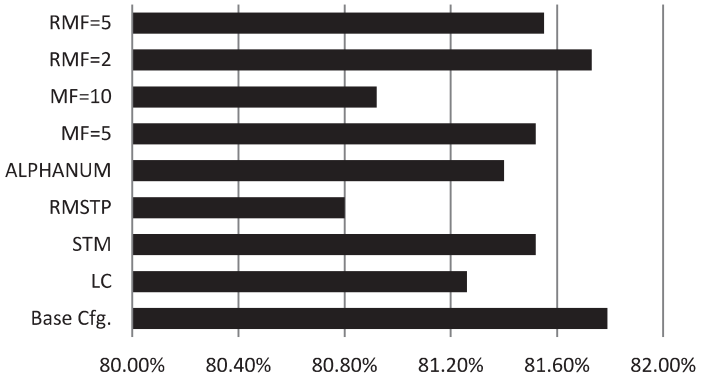

Figure 7.3 Performance of different configuration variations of word 1-gram feature at levels 1–3.

we first defined a baseline setting, which involved none of the normalization and filtering treatment mentioned in Section 7.3.1 We then applied one treatment at a time to observe its effect. Figure 7.3 presents the classification results of using each variation of word uni-gram features on the L2 proficiency group of levels 1–3 alone. The results in the other proficiency groups were similar.

We can see that the base configuration set achieved the highest accuracy among all variations. Any attempt to use less specific features degraded the performance. In particular, removing stop words resulted in the greatest drop of classification accuracy, which indicated that stop words were informative features for NLI. Nevertheless, setting the recurring minimum frequency to 2 (RMF $=$ 2) attained nearly the same results as that of base configuration set, while reducing the feature dimensions by more than a half (14,921 versus 36,012). This meant the treatment was useful: it discarded features that applied to only one individual author and did not generalize to his or her L1 language group, thus reducing noisy features. As a result, in subsequent experiments that involved word n-gram features or any other lexicalized features, we set the RMF as 2.

Figure 7.4 shows the classification accuracies of different character 5-grams for the L2 proficiency groups of levels 1–3. Similar trends were observed on other n-grams and L2 proficiency groups. Again, normalization and filtering treatment resulted in a slight decrease in the classification accuracy across different L2 proficiency groups. This meant that for NLI on the data set of EFCAMDAT, more specific features led to higher classification accuracy.

Meanwhile, the optimal length for word/character n-gram features varied across different proficiency levels. In general, a larger n was preferred as the

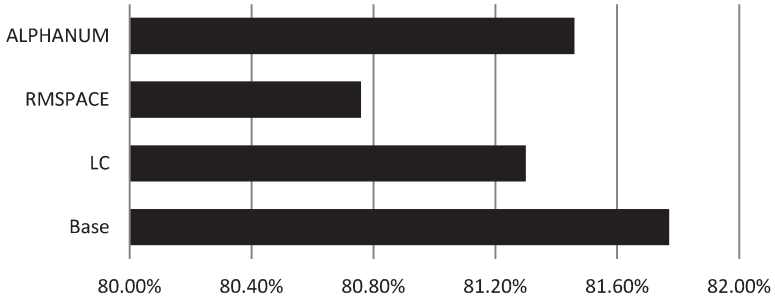

Figure 7.4 Performance of different variations of character 5-gram at levels 1–3.

proficiency level increased. This might be attributable to the fact that longer and more complex expressions were featured in the writings of advanced learners.

Figure 7.5 shows the NLI accuracies achieved by various POS features on the L2 proficiency group of levels 1–3. Again, the patterns on other L2 proficiency groups were similar. As we can see, the POS+Word subtype outperformed the other two subtypes. However, this advantage shrank as the order n grew. This was possibly because for lexicalized POS+Word features, uni-grams were already expressive and distinguishable, whereas higher-order n-grams suffered from data sparsity. On the other hand, for less specific POS features (pure POS and hybrid POS), higher-order POS sequences brought in more discriminatory information on native language backgrounds.

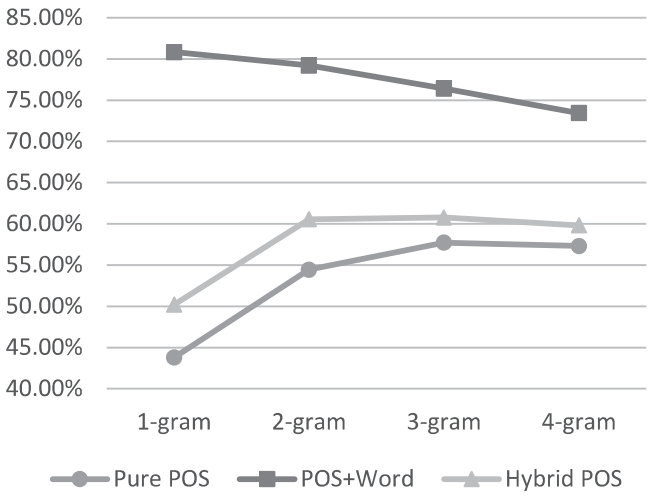

Figure 7.5 Performance of different POS n-grams at levels 1–3.

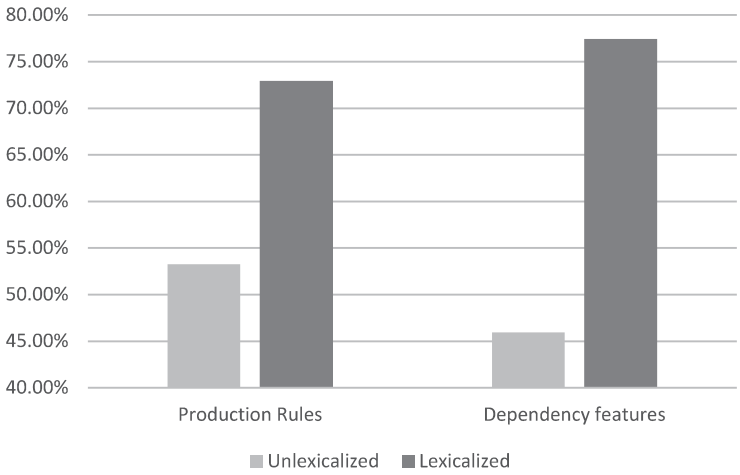

Figure 7.6 Performance of unlexicalized and lexicalized syntactic relational features at levels 1–3.

For production rules and dependencies, their lexicalized version also performed better than the unlexicalized one in most cases: Figure 7.6 demonstrates the classification accuracies of these two settings on the L2 proficiency group of levels 1–3. Similar results were found in other L2 proficiency groups.

Comparison across individual features This section compares the accuracy of NLI across individual features. Table 7.7 shows the accuracy of our NLI system when using each individual feature type alone. Here we report the results for the best configuration of the features only.

As shown in Table 7.7, lexical features (word and character n-grams) significantly outperformed syntactic ones (POS n-grams, production rules and dependencies) by up to 36%. Furthermore, the discriminatory effect of lexical features was more significant for beginners than for advanced learners. This might result from the fact that the former just started to learn morphology and

Table 7.7 *Performance of each individual feature type*

	1–3	4–7	8–11	12–16
Word	81.16%	79.17%	77.57%	63.12%
Char	81.19%	81.60%	79.30%	66.89%
POS	57.32%	62.37%	62.81%	56.29%
PR	51.88%	58.95%	60.32%	53.71%
Depd	45.16%	51.50%	55.06%	49.62%

lexicon, and were thus subject to more influence from L1. For syntactic features, the impact of L1 was clearer at medium than at low or high proficiency levels. This was probably because at the beginner levels, students were exposed to simple syntactic constructions that were relatively easy to learn, while by the time they got to the advanced levels, most students had a better grasp of grammar and their L1 background had less impact. POS tags were the most telling of the three syntactic features probably because they can tap into morpho-syntax as well as lexically driven patterns, whereas production rules and dependencies were too abstract.

7.4.2 Combination of Features

This section investigates the combined effect of the features in NLI. Table 7.8 shows the results of using all but one of the feature types. We can see that lexical features contributed to overall performance in almost all cases, with the only exception for character n-gram for levels 1–3. When combined with other feature types, syntactic features like production rules and dependencies played a less important role in NLI as the proficiency level of the students increased, whereas POS n-gram became more and more indispensable.

To conclude, we observe that most of the individual feature types had significant impact on the overall results. However, the impact of individual features changes across proficiency, indicating that the language of learners changes. It is also noticeable that the overall accuracy drops at the highest levels 12–16. This is probably due to the fact that, at most advanced levels, learner language is less influenced by L1 as learners acquire most of their target English L2. From a practical perspective, the results also suggest that when building NLI system for L2 data across different proficiency levels, we should consider different combination of features for different L2 proficiency levels.

Table 7.8 *Accuracy gain (+%) or loss (–%) of leave-one-out experiments*

Lvl	1–3	4–7	8–11	12–16
All	82.09%	82.54%	80.84%	69.50%
w/o Word	–0.66	–0.45	–0.37	–0.53
w/o Char	+0.68	–2.09	–2.11	–2.57
w/o POS	+0.73	+0.43	–0.42	–1.20
w/o PR	–0.18	–0.24	+0.48	+1.38
w/o Depd	+0.11	+0.24	+0.27	+0.40

Table 7.9 *The top 5 features for different feature types and proficiency levels. Please refer to the Penn Treebank POS tag set (Mitchell Marcus, 2012) and the Stanford parser (De Marneffe and Manning, 2008) for the meanings of POS tags, production rules and typed dependencies. Underscores "_" in character 6-grams represent a white space between words. COMMA and DOT refer to the punctuations of comma and dot.*

	Word n-grams (n = 1)
Lvl 1–3	*russia*; *brazil*; *china*; *moscow*; *paulo*
Lvl 12–16	*which*; *brazil*; *that*; *it*; *suitable*
	Char n-grams (n = 6)
Lvl 1–3	*Russia*; *m_Russ*; *China_*; *om_Bra*; *m_Bras*
Lvl 12–16	*which*; *brazil*; *becaus*; *As_for*; *_suita*
	POS n-grams (n = 2)
Lvl 1–3	*FW NNP*; *NNP FW*; *NNP NNP*; *MD VB*; *PRP MD*
Lvl 12–16	*COMMA PRP*; *COMMA IN*; *COMMA CC*; *NNS DOT*; *NN PRP*
	Production rules
Lvl 1–3	*NP→NNP+FW+NNP*; *VP→MD+VP*; *VP→MD+RB+VP*; *PP→IN+NP*; *ADJP→NP+JJ*
Lvl 12–16	*S→PP+NP+VP*; *S→CC+NP+VP*; *S→S+CC+S*; *S→SBAR+NP+VP*; *S→ADVP+NP+VP*
	Dependencies
Lvl 1–3	*neg*; *npadvmod*; *ccomp*; *det*; *prep_opposite*
Lvl 12–16	*prepc_as_for*; *prep_about*; *prep_of*; *prep_from*; *preconj*

7.4.3 Qualitative Analysis

We selected up to 100 best-performing features for each feature type using the Information Gain (IG) (Yang and Pedersen, 1997) criteria. IG measures the information (in the number of bits) obtained for classification by knowing the presence or absence of a feature. Table 7.9 shows the top five features for word uni-grams, character 6-grams, POS bi-grams, production rules, and dependencies, respectively for the lowest proficiency group and the highest proficiency group. The top features with other configurations (e.g., n-grams of different orders) had similar trends and are not shown here.

As shown in Table 7.9, the most indicative features varied a lot from one proficiency level to another. Take word n-grams, for example: the best-performing features for the beginners were country names that expressed one's L1 background explicitly. Proper names are replaced by function words such as *which* and *that* at higher levels, which indicates the growth of the learners' grammatical knowledge. We also notice a shift from phrasal rules such as *NP→NNP+FW+NNP* at lower levels to sentence level rules such as

Table 7.10 *Example of unique punctuations used in particular class: Brazilians(br) make more mistakes of replacing a quote mark with an acute accent mark*

Lvl	Type	Feature	ru	cn	br	Example
1–3	Char.	*'t*	4	0	**238**	"I don 't"
8–16	Char.	*'s*	3	0	**338**	"It 's"

$S \rightarrow PP+NP+VP$ at higher proficiency levels. These results suggested that features corresponding to more complex syntactic structures tended to be more informative for NLI on advanced learners.

We also performed a qualitative analysis of these representative features. Some of our findings are summarized in the following sections.

Punctuation and lexical preferences Results showed that Brazilians were more likely to mistake the acute accent mark ' (ASCII = 180) for the standard single quote '(ASCII = 47). This phenomenon remained in the writings by Brazilians of higher English proficiency level, as is shown in the following table.

Chinese learners did not use the dash mark as frequently as Russian and Brazilian learners did, which is shown in Table 7.11. This might be attributed to the native language transfer: since the dash mark was rarely used in the written text in Chinese, Chinese learners used it less in their English writings.

Some phrases were used more frequently by the authors of a particular native language background, such as those listed in Table 7.12. These phrases were manually checked to ensure that they distributed across different topics and across different proficiency levels. They did occur more frequently in the writings by Russian or Chinese learners.

Clause-initial prepositional phrase (CIPP) A syntactic feature discriminating between Russians, Brazilians, and Chinese was the production rule PP-NP-VP, which involved sentences in which a prepositional phrase (PP) appears at the beginning of a clause as in (1).

Table 7.11 *Chinese Learners tend to underuse dash*

Lvl	Type	Feature	ru	cn	br
1–3	Word	–	2327	**64**	1535
8–16	Char.	–	1130	**93**	843

Table 7.12 *Example of phrases frequently used by particular class*

Lvl	Type	Feature	ru	cn	br	Example
8–16	Word	*as for me*	**127**	5	0	"As for me I prefer to eat at home"
8–16	Word	*to my mind*	**80**	1	1	"To my mind it is the most important thing for me."
4–7	Word	*try my best*	1	105	0	"I will try my best"
4–7	Word	*so I*	595	**1987**	804	"So I decide to leave"
8–16	Word	*whats more*	6	**89**	0	"Whats more, there are too much oil of main course."
8–16	Word	*have a try*	4	**77**	0	"I urge you to have a try"

(1) 1 *in 18:00 o'clock* he has dinner
2 *In the evening* he eats dinner at 18 o'clock
3 *according to market research* our clients consider that our logo is old fashioned
4 *opposite the window* there is a big TV

What was interesting about this feature was that the direction of correlation with national language changed across proficiency. At the early beginner levels 1–3, Chinese learners produced clause initial PPs much more often than Russians and Brazilians did: "ru": 725, "cn": 1672, "br": 849. However, in late beginner/early intermediate levels 4–7, it was the Russians that were the most productive, followed by Brazilians: "ru": 3626, "cn": 1397, "br": 2798. This trend seemed consolidated at intermediate and advanced levels 8–16: "ru": 1584, "cn": 247, "br": 776.

It was reasonable to hypothesize that the shift in the use of this rule from early beginner levels was linked to qualitative changes in the way this rule was used. We therefore inspected the actual productions to gain some insight. We compared productions for early beginner levels, when the rule was used dominantly by Chinese learners, with productions from all the other levels when Russians and Brazilians used this rule more.

CIPP at Levels 1–3

At this level learners of all nationalities used clause-initial prepositional phrases to convey primarily temporal information, like the time of day and the day of week etc. as in (2).

(2) 1 *In the afternoon* he goes shopping at 3 o'clock (**cn**)
2 *On Sunday*, he goes to the park and meets friends, and *at half past eleven* he palys tennis with his friends (**cn**)
3 *On Saturday at eleven-thirty* he was going to swim (**ru**)
4 *In the evening at 6 o'clock* he eat dinner (**ru**)

But there were some differences in the phrases used by different learners. First of all, Russians tended to use more complex prepositional phrases with two points of time reference, e.g., day of the week and time as in (2)c-d. This pattern was not as productive with Chinese and Brazilian learners. Meanwhile, a fact that set Russian learners apart from Brazilians and Chinese was the production of a wider range of phrases that went beyond strict reference to time and location as indicated in the following examples.

(3) 1 I think we have to say that *in success case* they will get additional bonuses (**ru**)
2 I want to learn English that *in future* I'll can understand other people (**ru**)
3 *With great pleasure* we inform out clients and shareholders of the change of company's logo (**ru**)
4 so *in the nearest time* we do not expect good growth (**ru**)
5 Workspaces are not clean and tidy : *in fact* they are messy (**ru**)

Second, Brazilians introduced many of their temporal phrases with *after* as in (4). In examples like (4) the temporal prepositional phrase established a sequential link between events (rather than pointing to a specific point in time).

(4) 1 and *after breakfast* I go to work with my dad (**br**)
2 *after the movie* We have some coffee (**br**)

Meanwhile, Brazilians used prepositional phrases to mark not only time but also space/location as in (5). While we can find such uses in the productions of Chinese and Russian learners, Brazilians appeared much more likely to bring information on location at the beginning of their sentences.

(5) 1 *On Park Road* there is a movie theater, *near the movie theater* there are many clothes stores and books stores (**br**)
2 and *in coffe room* there is one table (**br**)
3 In opposite you have one pharmacy, and *on the left* you have the a market (**br**)
4 There are two big windows and *opposite my bed* I put a table with a TV on it (**br**)

CIPP at Levels 4–16

As expected, learners used a wider range of prepositional phrases in higher proficiency levels, a fact reflecting their ability to better express themselves in their L2 English. The range of clause-initial prepositional phrases was extended for all three nationalities, as shown in (7).

(6) 1 I can not go surfing because *for me* it is too scary (**ru**)
2 and *in the process* she changes size several times (**ru**)
3 but *at last* these girls decided not to attend (**cn**)

Reference to time and location remains dominant, but it involves more complex or abstract temporal meanings as indicated in (7).

(7) 1 She set the conditions that *before the wedding* she will be traveling on the ship (**ru**)
2 I hope that *after all my efforts* I'll get a job at the company of my dream (**ru**)
3 so *until that time* it's sure of that I will share more time with my wife and my baby (**cn**)
4 but *at the end* everything was resolved (**br**)
5 I hope that *during my absence* you take care of everything (**br**)

The observations mentioned earlier indicate that our methodology can capture changing patterns in learner language across proficiency. To explain these changing patterns, we hypothesize that the overuse of the rule by Chinese learners at early beginner stages compensates for the absence of tense morphology in their grammar, an area that is known to be challenging for Chinese learners (Lardiere, 1998). A prediction of this hypothesis is that the drop in the use of PP preposing should correlate with increasing use and accuracy of verbal tense morphology. Regarding Russians, we hypothesize that preposing is due to transfer from L1 Russian where preposing is frequent for information structure (King, 1995). A prediction of this hypothesis is that the increase in PP preposing in Russian learners correlates with productive marking of information structure (e.g., higher accuracy in nominal anaphora, etc). Testing these predictions is beyond the scope of this chapter. However, the more general point is that NLI in a longitudinal corpus can capture both L1 effects as well as changing patterns across the learning trajectory, which can lead to new hypotheses for researchers in second language acquisition.

7.5 Conclusion

We developed a method for NLI that employs accurate machine learning (SVM) with a wide range of linguistic features (ranging from character features to syntactic dependencies) and applied this method to the newly developed, large EFCAMDAT corpus that, unlike previous learner corpora, provides longitudinal data at multiple proficiency levels. For the first time, we compared the performance of different feature types in NLI at different proficiency levels. We reported high overall accuracy of around 80% at low and medium proficiency levels and 70% at advanced levels. Our quantitative and a qualitative analysis of different features revealed that the top performing features differed from one proficiency level to another. Our linguistic analysis showed that our results can be of interest to research on L2 acquisition.

In the future, we plan to investigate NLI at finer-grained levels of proficiency and to integrate a wider range of nationalities, exploring strategies to deal with data sparsity. We also plan to develop new NLI methodology suitable for the analysis of large, longitudinal data, based on the insights gained in our experiments. Finally, we plan to conduct further linguistic evaluation of the data.

Acknowledgments

We thank EF Education First for providing the data, sponsoring the development of EFCAMDAT and the EF Research Lab for Applied Language Learning, the Isaac Newton Trust (Cambridge) for a grant supporting the development of EFCAMDAT, and finally the Royal Society, UK.

References

Ahn, Charles S. (2011). "Automatically detecting authors' native language". Ph.D. thesis, Monterey, California. Naval Postgraduate School.

Al-Rfou, Rami. (2012). "Detecting English Writing Styles For Non-native Speakers". In: *arXiv preprint arXiv:1211.0498*.

Bestgen, Yves, Sylviane Granger, Jennifer Thewissen, et al. (2012). "Error patterns and automatic L1 identification". In: *Approaching language transfer through text classification*, pp. 127–153.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, pp. 338–339.

Blanchard, Daniel, et al. (2013). "TOEFL11: A corpus of non-native English". In: *Educational Testing Service*.

Brooke, Julian, and Graeme Hirst (2011). "Native language detection with cheap learner corpora. In: *Conference of Learner Corpus Research (LCR2011)*.

Bykh, Serhiy, and Detmar Meurers (2012). "Native Language Identification Using Recurring N-grams–Investigating Abstraction and Domain Dependence". In: *Proceedings of COLING 2012: Technical Papers*, pp. 425–440.

Charniak, Eugene, and Johnson, Mark. (2005). "Coarse-to-fine n-best parsing and MaxEnt discriminative reranking". In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 173–180.

De Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D Manning (2006). "Generating typed dependency parses from phrase structure parses". In: *Proceedings of LREC*, Vol. 6, pp. 449–454.

De Marneffe, Marie-Catherine, and Christopher D Manning (2008). "Stanford typed dependencies manual". In: URL http://nlp.stanford.edu/software/dependenciesmanual.pdf.

Estival, Dominique et al. (2007). " Author profiling for English emails". In: *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING07)*, pp. 263–272.

8 Evaluating Language Acquisition Models: A Utility-Based Look at Bayesian Segmentation

Lisa Pearl and Lawrence Phillips

Abstract

Computational models of language acquisition often face evaluation issues associated with unsupervised machine learning approaches. These acquisition models are typically meant to capture how children solve language acquisition tasks without relying on explicit feedback, making them similar to other unsupervised learning models. Evaluation issues include uncertainty about the exact form of the target linguistic knowledge, which is exacerbated by a lack of empirical evidence about children's knowledge at different stages of development. Put simply, a model's output may be good enough even if it does not match adult knowledge because children's output at various stages of development *also* may not match adult knowledge. However, it is not easy to determine what counts as "good enough" model output. We consider this problem using the case study of speech segmentation modeling, where the acquisition task is to segment a fluent stream of speech into useful units like words. We focus on a particular Bayesian segmentation strategy previously shown to perform well on English, and discuss several options for assessing whether a segmentation model's output is good enough, including cross-linguistic utility, the presence of reasonable errors, and downstream evaluation. Our findings highlight the utility of considering multiple metrics for segmentation success, which is likely also true for language acquisition modeling more generally.

8.1 Introduction

A core issue in machine learning is how to evaluate unsupervised learning approaches (von Luxburg, Williamson, & Guyon, 2011), since there is no a priori correct answer the way that there is for supervised learning approaches. Computational models of language acquisition commonly face this problem because they attempt to capture how children solve language acquisition

tasks without explicit feedback, and so typically use unsupervised learning approaches. Moreover, evaluation is made more difficult by uncertainty about the exact nature of the target linguistic knowledge and a lack of empirical evidence about children's knowledge at specific stages in development. Given this, how do we know that a model's output is "good enough"? How should success be measured? To create informative cognitive models of acquisition that offer insight into how children acquire language, we should consider how to evaluate acquisition models appropriately (Pearl, 2014; Phillips, 2015; Phillips & Pearl, 2015b).

As a case study, we investigate the initial stages of speech segmentation in infants, where a fluent stream of speech is divided into useful units, such as words. For example, the acoustic signal transcribed via IPA as /ajlʌvðizpɛŋgwɪnz/ (*Ilovethesepenguins*) might be segmented as /aj lʌv ðiz pɛngwɪnz/ (*I love these penguins*). A particular Bayesian segmentation strategy has been shown to be quite successful on English (Goldwater, Griffiths, & Johnson, 2009; Pearl, Goldwater, & Steyvers, 2011; Phillips & Pearl, 2012, 2014a, 2014b; Phillips, 2015; Phillips & Pearl, 2015b), particularly when cognitive plausibility considerations have been incorporated at both the computational and algorithmic levels of Marr (1982). One way to evaluate if this strategy is "good enough" is to see how it fares cross-linguistically. This is based on the premise that core aspects of the language acquisition process – such as the early stages of segmentation occurring in six- to seven-month-olds (Thiessen & Saffran, 2003; Bortfeld, Morgan, Golinkoff, & Rathbun, 2005) – are universal. So, a viable learning strategy for early segmentation should succeed on any language infants encounter.

Traditionally, a segmentation model's output has been compared against a "gold standard" derived from adult orthographic segmentation (e.g., Brent, 1999; M. Johnson, 2008; Goldwater et al., 2009; Blanchard, Heinz, & Golinkoff, 2010; M. Johnson, Demuth, Jones, & Black, 2010; M. Johnson & Demuth, 2010; Pearl et al., 2011; Lignos, 2012; Fourtassi, Börschinger, Johnson, & Dupoux, 2013). Notably, orthographic segmentation assumes the desired units are orthographic words. However, if we look at the world's languages, it becomes clear that it is also useful to identify morphemes – the smallest meaningful linguistic units – particularly for languages with regular morphology. That is, infants might reasonably segment morphemes from fluent speech rather than entire words. Notably, models that identify subword morphology are penalized by the gold standard evaluation, and this highlights the need for a more flexible metric of segmentation performance. More generally, a segmentation strategy that identifies units useful for later linguistic analysis should not be penalized.

Still, how do we know that the segmented units are truly useful? If we believe that the output of early segmentation scaffolds later acquisition

processes, a useful segmentation output should simply enable these later processes to successfully occur (Phillips & Pearl, 2015a). For example, one goal of early segmentation is to generate a proto-lexicon in order to bootstrap language-specific segmentation cues like stress pattern (e.g., in English, words in child-directed speech tend to begin with stress [Swingley, 2005], and the same is true for child-directed German and Hungarian [Phillips & Pearl, 2015a]). Does the inferred proto-lexicon of units yield the appropriate language-specific cue? As another example, infants begin to learn mappings from word forms to familiar objects as early as six months (Tincoff & Jusczyk, 1999; Bergelson & Swingley, 2012; Tincoff & Jusczyk, 2012). Can the inferred proto-lexicon be used successfully for this process?

In the remainder of this chapter, we first review relevant aspects of infant speech segmentation, including what is known about the developmental trajectory, the cues infants are sensitive to, and how infants perceive the input. This forms the empirical basis for the modeled Bayesian segmentation strategy, which we then discuss in terms of its underlying generative assumptions and the algorithms used to carry out its inference. We turn then to the cross-linguistic evaluation of this strategy over input derived from child-directed speech corpora in seven languages from the CHILDES database (MacWhinney, 2000): English, German, Spanish, Italian, Farsi, Hungarian, and Japanese. This section includes comparison to the gold standard as well as a more flexible metric derived from the gold standard that's consistent with children's imperfect early segmentation behavior. We find that the Bayesian strategy seems to be "good enough" cross-linguistically. This is especially true once we use this more nuanced output evaluation that considers potentially useful nonword units valid. This serves as a general methodological contribution about the definition of segmentation success, especially when we consider that useful units may vary across the world's languages.

We conclude with an evaluation metric that is even more utility-driven: whether the output of the segmentation strategy is helpful for subsequent acquisition processes, such as inferring a language-specific stress-based segmentation cue and learning early word-meaning mappings. Interestingly, just because a strategy yields more accurate segmentations when compared against the gold standard does not mean it is always more useful for subsequent acquisition processes. This underscores the value of considering multiple metrics for segmentation success, in addition to the traditional comparison against the gold standard.

8.2 Early Speech Segmentation

Segmentation is not easy – words blur against one another, making speech more like a stream of sound rather than something divided into discrete,

separable chunks (Cole & Jakimik, 1980). Yet, speech segmentation is one of the first tasks infants accomplish in their native language, and the resulting segmented units underlie subsequent processes such as learning word meanings, syntactic categories, and syntactic structure. In order to accurately model the segmentation process, we need to know the empirical data that form the basis for decisions regarding the model's learning assumptions, input, inference, and evaluation.

8.2.1 When Does Early Speech Segmentation Begin?

The first behavioral evidence for speech segmentation in infants comes at six months (Bortfeld et al., 2005), when infants seem to know a small set of very frequent words (Bergelson & Swingley, 2012). They recognize these words in speech and can use them to segment new utterances. Between seven and nine months, infants learn to utilize language-specific cues such as stress pattern (Jusczyk, Cutler, & Redanz, 1993; Jusczyk, Houston, & Newsome, 1999; Thiessen & Saffran, 2003), phonotactics (Mattys, Jusczyk, & Luce, 1999), allophonic variation (Hohne & Jusczyk, 1994; Jusczyk, Hohne, & Baumann, 1999), and coarticulation (E. Johnson & Jusczyk, 2001). These language-specific cues are typically more reliable than language-independent cues such as transitional probability between syllables. However, it turns out that infants around seven months old prefer to rely on transitional probability information alone rather than language-specific cues such as stress patterns (Thiessen & Saffran, 2003), even though transitional probability is a less reliable cue. Neuroimaging evidence from neonates suggests that this sensitivity to statistical cues like transitional probabilities is present at birth (Teinonen, Fellman, Näätänen, Alku, & Huotilainen, 2009). This in turn suggests that the initial stages of speech segmentation rely on cues that are independent of the particular language being segmented (e.g., the process of tracking transitional probabilities does not vary from language to language, though the probabilities themselves clearly do). It is only after this initial stage that infants harness the more powerful language-dependent cues that do vary between languages (e.g., the specific stress-based segmentation cues that differ between English and French). So, a model of early speech segmentation should also likely rely only on language-independent cues.

8.2.2 The Unit of Infant Speech Perception

The basic unit of infant speech perception has been a source of significant debate for some time (see Phillips, 2015 and Jusczyk, 1997 for a more detailed review of this debate). Experimental studies have typically focused on whether the basic representational unit for infants is syllabic or segmental (Jusczyk

& Derrah, 1987; Bertonicini, Bijeljac-Babic, Jusczyk, Kennedy, & Mehler, 1988; Bijeljac-Babic, Bertoncini, & Mehler, 1993; Jusczyk, Jusczyk, Kennedy, Schomberg, & Koenig, 1995; Eimas, 1999). Jusczyk (1997, p. 115) summarizes several studies by noting that "there is no indication that infants under six months of age represent utterances as strings of phonetic segments." Instead, evidence for segmental representations of speech is mostly present in infants older than six months: infants first begin to ignore vowel contrasts that aren't relevant for their native language around six months (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Polka & Werker, 1994), while irrelevant consonant contrasts are ignored between eight and twelve months (Werker & Tees, 1984; Werker & Lalonde, 1988; Best, McRoberts, & Sithole, 1988; Best, McRoberts, LaFleur, & Silver-Isenstadt, 1995).

More generally, as infants get older, they are better able to represent information about segments; in contrast, they appear relatively comfortable with syllables from early on. This can be seen in infants' ability to track statistical relationships: while transitional probabilities over syllables can be tracked at birth (Teinonen et al., 2009), transitional probabilities over segments first seem to occur around nine months (Mattys et al., 1999). Given this, a reasonable assumption for a model meant to capture segmentation strategies being used by six-month-olds would be that the input is perceived as a stream of syllables.

8.2.3 Constraints on Infant Inference

One thing that makes child language acquisition so impressive is that it is accomplished despite the many cognitive constraints imposed by the developing brain. Though there has been increasing interest in cognitively constrained language acquisition models (e.g., Anderson, 1990; Shi, Griffiths, Feldman, & Sanborn, 2010; Bonawitz, Denison, Chen, Gopnik, & Griffiths, 2011; Pearl et al., 2011; Phillips & Pearl, 2015b), there isn't very much experimental evidence to suggest exactly what kinds of constraints should be imposed. There are many possibilities, but we focus on three that have been incorporated into past acquisition models and that seem reasonable starting points: online processing, nonoptimal decision-making, and recency effects.

Online processing refers to the idea that data are processed as they are encountered, rather than being stored in explicit detail for later batch processing. It is generally accepted that this is a reasonable constraint for human language processing, and commonly used as justification that a model operates at the algorithmic level in the sense of Marr (1982), rather than being idealized (e.g., Lignos & Yang, 2010; Pearl et al., 2011; Lignos, 2012; Phillips & Pearl, 2014b, 2014a, 2015b). So, this is likely reasonable to incorporate into infant inference.

For decision-making, experimental evidence from infants and children suggest that they do not always choose the highest probability option available, which would be considered the optimal decision (Köpcke, 1998; C. H. Kam & Newport, 2005; C. L. H. Kam & Newport, 2009; Davis, Newport, & Aslin, 2011; Denison, Bonawitz, Gopnik, & Griffiths, 2013). Instead, children sometimes appear to probabilistically sample the available options (Davis et al., 2011; Denison et al., 2013). Other times, they appear to simply generalize to a single option, which might in fact be a lower probability option (Köpcke, 1998; C. H. Kam & Newport, 2005; C. L. H. Kam & Newport, 2009). This suggests that infant inference may involve nonoptimal decision-making.

With respect to memory constraints, experimental evidence suggests that a recency bias exists in infants (Cornell & Bergstrom, 1983; Gulya, Rovee-Collier, Galluccio, & Wilk, 1998; Rose, Feldman, & Jankowski, 2001), where the most recently encountered data have privileged status. So, this is also reasonable to incorporate into infant inference.

8.2.4 What We Know about Segmentation Output

As mentioned before, one empirical checkpoint for segmentation is that a strategy ought to be successful for any human language. Beyond that, we also have some evidence about the kinds of errors children produce – this underscores that successful early segmentation does not necessarily mean adultlike segmentation. For example, Brown (1973) and Peters (1983) find that even three-year-old children still produce missegmentations. These errors can be broadly split into two types: function word undersegmentations (e.g., *that'sa*, *it'sa*) and function word oversegmentations (e.g., *a nother*, *be have*). So, segmentation error patterns may provide a useful qualitative benchmark for model output, and have been previously used this way (Lignos, 2012; Phillips & Pearl, 2012, 2015b).

8.3 A Bayesian Segmentation Strategy

Bayesian segmentation strategies combine the prior probability of a potential segmentation s for an utterance u with the likelihood in order to generate the posterior probability of s ($P(s|u)$) using Bayes' rule, as shown in (8.1).

$$P(s|u) \propto P(s)P(u|s) \tag{8.1}$$

The Bayesian strategy we investigate builds off of two fundamental insights about the infant's inferred proto-lexicon, both of which were used in an earlier segmentation strategy by Brent (1999). First, frequent words should be preferred over infrequent words. Second, shorter words should be preferred over longer words. These parsimony biases were incorporated by Goldwater et al.

(2009) into a Bayesian strategy that used a Dirichlet Process (Ferguson, 1973) to determine the prior probability of a segmentation.

The Dirichlet Process (DP) is a nonparametric stochastic process resulting in a probability distribution often used in Bayesian modeling as a prior because it has properties well suited to language modeling. First, because it is nonparametric, it does not need to prespecify the number of items (e.g., word types) that might be encountered. Second, the DP facilitates "rich-get-richer" behavior, where frequent items are more likely to be encountered later. This is useful because word frequencies in natural languages tend to follow a power-law distribution that the DP naturally reproduces due to this behavior (Goldwater, Griffiths, & Johnson, 2011).

Goldwater et al. (2009) implemented two versions of the DP segmentation strategy that differed in their generative assumptions. Both versions use the likelihood function, i.e., the probability of an utterance given its segmentation $P(u|s)$, to simply rule out potential segmentations that do not match the observed utterance. For example, a possible segmentation /ðə pɛŋgwɪn/ (*the penguin*) does not match an observed speech stream /ðəkɪɾi/ (*thekitty*) when the possible segmentation is concatenated (*thepenguin≠thekitty*). So, this possible segmentation would have a likelihood of 0. In contrast, the possible segmentation /ðəki ɾi/ (*theki tty*) does match (*thekitty=thekitty*), and so would have a likelihood of 1. Where the DP strategy versions differ is how the prior probability for a segmentation is determined. We describe each version in turn before reviewing the inference algorithms paired with each generative model.

8.3.1 DP Segmentation: Unigram Assumption

The first version of the DP segmentation strategy uses a unigram language model (DP-Uni), with the modeled learner naively assuming that each word is chosen independently of the words around it. The prior of the potential segmentation is calculated using this generative assumption. To do this, the model must define the probability of every word in the utterance, and so a DP-Uni learner assumes that for any utterance, each segmented word w_i is generated by the following process:

1. If w_i is not a novel lexical item, choose an existing form ℓ for w_i.
2. If w_i is a novel lexical item, generate a form (e.g., the syllables $x_1 \ldots x_M$) for w_i.

Because the model does not know whether w_i is novel, it has to consider both options when calculating the probability of the word. We note that deciding whether w_i is novel is not the same as deciding whether the form of w_i has been previously encountered. As an example, consider the first time the modeled learner encounters the sequence /et/ (such as in the word *ate*). Because

$count_{/et/} = 0$, the word must be novel ($count_{/et/}$ then $= 1$). Now, suppose the same sequence is encountered again, but from the word *eight*. The learner must decide if this sound sequence is a second instance of the /et/ it saw before (*ate*) or instead the first instance of a novel /et/ type (such as in the word *eight*). In the first case, the count might be updated to $count_{/et/} = 2$; in the second case, the counts might be updated to $count_{/et/} = 1$ and $count_{/et/_2} = 1$. This particular aspect of the DP distribution naturally allows for the existence of homophones (e.g., *ate*/*eight*) without requiring any additional machinery.

Returning to the DP generation process, generating either non-novel or novel items is fundamental to the DP. In a classic DP, the probability of generating a non-novel item is proportional to the number of times that item has been previously encountered. This is shown in (8.2), where n_ℓ refers to the number of times lexicon item ℓ has been seen in the set of words previously encountered, denoted as w_{-i}. In the denominator, i represents the total number of words encountered thus far, including the word previously under consideration. Because the current word is not included in n_ℓ, 1 is subtracted from it.

$$P(w_i = \ell, w_i \neq novel | w_{-i}) = \frac{n_\ell}{i - 1 + \alpha} \tag{8.2}$$

Equation (8.2) gives higher probability to word types that have been encountered before. So, the more a word type is encountered by the modeled learner, the more often the modeled learner will prefer it in the future. This will end up generating the power-law frequency distribution found in natural languages.

When the DP instead generates a novel word, the word is not represented as an atomic whole, but rather constructed from its individual parts, such as syllables. To model this, we make use of the P_0 in (8.3) to describe the probability that any word might be made up of a particular string of subword units $x_1 \ldots x_M$. Each subword unit x_j is generated in turn for all M units in the word, with the probability of x_j treated as a uniform choice over all possible subword units in the corpus.[1]

$$P_0(w_i = x_1 \ldots x_M) \propto \prod_{j=1}^{M} P(x_j) \tag{8.3}$$

The probability of generating a novel item in a DP is weighted by the free model parameter α, also known as the DP concentration parameter. This parameter has an intuitive interpretation, where higher values of α lead to a preference

[1] The model additionally includes the generation of word and utterance boundaries, with a word ending with some probability $p_\#$ and an utterance ending with some probability $p_\$$. See Goldwater et al. (2009) for discussion of these model components in the unigram and bigram versions of this strategy.

for generating novel words in the proto-lexicon.[2] The full probability of generating a novel word is therefore described by equation (8.4).

$$P(w_i = \ell, w_i = novel|w_{-i}) = \frac{\alpha P_0(w_i = x_1 \ldots x_M)}{i - 1 + \alpha} \tag{8.4}$$

Both equations 8.2 and 8.4 can be combined to generate the full probability of a word being produced either as a non-novel or novel item.

$$P(w_i = \ell|w_{-i}) = \frac{n_\ell + \alpha P_0(w_i = x_1 \ldots x_M)}{i - 1 + \alpha} \tag{8.5}$$

8.3.2 DP Segmentation: Bigram Assumption

The second version of the DP segmentation strategy uses a bigram language model (DP-Bi), with the learner assuming (slightly less naively) that each word is chosen based on the word preceding it. Goldwater et al. (2009) model this using a hierarchical Dirichlet Process (Teh, Jordan, Beal, & Blei, 2006), with the generative process selecting bigrams, words, and subword units as follows:

1. If the pair $< w_{i-1}, w_i >$ is not a novel bigram, choose an existing form ℓ for w_i from those that have been previously generated after w_{i-1}.
2. If the pair $< w_{i-1}, w_i >$ is a novel bigram:
   If w_i is not a novel lexical item, choose an existing form ℓ for w_i.
   If w_i is a novel lexical item, generate a form $(x_1 \ldots x_M)$ for w_i.

As with the DP-Uni model, the DP-Bi model must make a decision between an item being novel or not; the main difference is that the DP-Bi model considers bigrams first. If a bigram is not novel, the DP-Bi model gives higher probability to bigrams that have been encountered before. If a bigram is instead novel, then the individual lexical item (the second word in the bigram) must also be generated. This is done with a DP in the same fashion as the unigram DP, making this a hierarchical DP. The probability of any bigram $< w_{i-1}, w_i >$ is then determined using equations (8.6), (8.7), and (8.3).

$$P(< w_{i-1}, w_i = \ell > |w_{-i}) = \frac{n_{< w_{i-1}, w_i = \ell >} + \beta P_1(w_i = \ell|w_{-i})}{n_{w_{i-1}} + \beta} \tag{8.6}$$

Equation (8.6) calculates the probability of the bigram $< w_{i-1}, w_i >$, given that the second word of the bigram w_i takes the form ℓ and considering all the words observed previously except w_i, denoted by w_{-i}. This includes the number of times ℓ appears as the second word of bigrams beginning with word w_{i-1} ($n_{< w_{i-1}, w_i = \ell >}$), as well as the total number of bigrams beginning with w_{i-1},

[2] A more thorough treatment of the role of various model parameters for both the unigram and bigram DP segmentation models can be found in Goldwater et al. (2009).

denoted by $n_{w_{i-1}}$. The concentration parameter β determines how often a novel second word is expected, with higher values indicating a general preference for more novel bigrams.

Equation (8.7) describes the process for generating a novel second word in the bigram.

$$P_1(w_i = \ell | w_{-i}) = \frac{t_{w_i=\ell} + \gamma P_0(w_i = x_1 \ldots x_M)}{t + \gamma} \tag{8.7}$$

A novel second word with form ℓ is based on the number of times any bigram with second word ℓ has been generated, $t_{w_i=\ell}$. The total number of novel bigrams is represented by t. The concentration parameter γ determines how often this novel second word is itself a novel lexical item, constructed from its constituent subword units $x_1 \ldots x_M$ using P_0, as in Equation (8.3).

8.3.3 DP Segmentation Inference

Pearl et al. (2011) used a variety of inference algorithms with these two DP segmentation strategies, including both idealized and constrained procedures. Idealized inference procedures provide a computational-level analysis in the sense of Marr (1982), and offer a best-case scenario of how useful the learning assumptions of the model are. Constrained inference procedures provide a more algorithmic-level analysis in the sense of Marr (1982). They in turn offer a more cognitively plausible assessment of how useable the learning assumptions are by humans, who have cognitive limitations on their inference capabilities (particularly infants). Here we focus on one inference algorithm of each kind.

8.3.3.1 Idealized Inference The original inference algorithm used by Goldwater et al. (2009) for DP segmentation was Gibbs sampling (Geman & Geman, 1984), a batch procedure commonly used for idealized inference, due to its guaranteed convergence on the optimal solution given the model constraints. Gibbs sampling initializes the model parameters (in this case potential word boundaries) and then updates each parameter value one at a time, conditioned on the current value of all other parameters. This process is repeated for a number of iterations until convergence is reached (e.g., Goldwater et al., 2009 and Pearl et al., 2011 used 20,000 iterations).

For DP segmentation, each possible boundary location is a parameter that either has a boundary or does not. Boundaries are initialized randomly, and the inference procedure goes through each boundary location in the corpus, deciding whether to place/remove a boundary, given the other current boundary locations. In particular, for each possible boundary, there is a choice between creating a single word (H_0) or two words (H_1) out of the two adjoining pieces. For example, H_0 might be /ðəkɪɾi/ (*thekitty*) while H_1 is /ðə kɪɾi/ (*the kitty*)

for the potential boundary location between /ðə/ and /kɪɾi/. The probability of inserting a boundary (H_1) can be defined as the normalized probability of H_1, shown in (8.8), with $P(H_0)$ and $P(H_1)$ defined by the DP-Uni or DP-Bi generative models:

$$Normalized\ P(H_1) = \frac{P(H_1)}{P(H_1) + P(H_0)} \tag{8.8}$$

The inference procedure then probabilistically selects either H_0 or H_1, based on their normalized probabilities. If no boundary is placed (H_0), only a single word has to be generated; if a boundary is placed (H_1), two words have to be generated. Intuitively, the model may prefer either H_0 because it requires fewer words or H_1 because it requires shorter words. The exact trade-off depends on the model parameters and, most importantly, on the frequency each word (or bigram) is currently perceived to have.

8.3.3.2 Constrained Inference Pearl et al. (2011) described several inference procedures that incorporate one or more of the cognitive limitations relevant for infant speech segmentation mentioned before: (1) online processing, (2) nonoptimal decision-making, and (3) a recency bias. We focus on the one that incorporates all three of these constraints to some degree (called the DMCMC constrained learner by Pearl et al., 2011). This inference procedure performs inference online, segmenting each utterance as it is encountered. It can also be thought to involve nonoptimal decision-making because it probabilistically samples whether to insert a boundary rather than always selecting the highest probability option.[3]

It additionally uses the Decayed Markov Chain Monte Carlo method (Marthi, Pasula, Russell, & Peres, 2002) to implement a recency bias. In particular, similar to the idealized inference procedure, it samples individual boundary locations and updates them conditioned on the value of all other currently encountered potential boundaries. However, instead of sampling all currently known potential boundaries equally, the locations to sample are selected based on a decaying function anchored from the most recently encountered potential boundary location (at the end of the current utterance). The probability of sampling potential boundary b_a, which is a potential boundaries away from the end of the current utterance, is determined by (8.9):

$$P(b_a) = \frac{a^{-d}}{\sum a_i^{-d}} \tag{8.9}$$

[3] Unlike Gibbs sampling, which also probabilistically samples whether to insert a boundary, there is no guarantee of convergence on the optimal solution.

Table 8.1 *Summary of the syllabified child-directed speech corpora, including the CHILDES database corpora they are drawn from (Corpora), the age ranges of the children they are directed at (Age range), the number of utterances (# Utt), the number of unique syllables (# Syl types), the average number of syllables per utterance (Syls/Utt), and the probability of a word boundary appearing between syllables (B Prob).*

Language	Corpora	Age range	# Utt	# Syl types	Syls/Utt	B Prob
English	Brent	0;6–0;9	28391	2330	4.16	76.26
German	Caroline	0;10–4;3	9378	1682	5.30	68.60
Spanish	JacksonThal	0;10–1;8	16924	522	4.80	53.93
Italian	Gervain	1;0–3;4	10473	1158	8.78	49.94
Farsi	Family, Samadi	1;8–5;2	31657	2008	6.98	43.80
Hungarian	Gervain	1;11–2;11	15208	3029	6.30	51.19
Japanese	Noji, Miyata, Ishii	0;2–1;8	12246	526	4.20	44.12

The parameter d implements the recency effect: larger values of d indicate stronger biases, concentrating the boundary sampling efforts on more recently encountered data. We discuss results using $d = 1.5$, which implements a strong recency bias: using this d value, Phillips and Pearl (2015b) found that 83.6% of sampled boundaries occurred in the current utterance in their corpus of English child-directed speech, 11.8% in the previous utterance, and only 4.6% in any other previous utterance.

8.4 How Well Does This Work Cross-Linguistically?

8.4.1 Cross-Linguistic Corpora

Phillips and Pearl (2014a, 2014b) evaluated the DP segmentation strategy on seven languages: English, German, Spanish, Italian, Farsi, Hungarian, and Japanese. These languages vary in many ways, including their morphology: some are more agglutinative and have more regular morphology systems (Hungarian, Japanese) while the others are fusional to different degrees and have less regular morphological systems (English, German, Spanish, Italian, and Farsi). Syllabified child-directed speech corpora were derived from the CHILDES database (MacWhinney, 2000; Gervain & Erra, 2012; Phillips & Pearl, 2015b),[4] and relevant summary statistics for them are shown in Table 8.1.

We can make a few observations. First, not all languages had corpora available of speech directed at children younger than a year old, so the age range does vary. Second, the corpora vary in size, though this does not appear to

[4] See Phillips (2015) for details of this process.

negatively impact the results for smaller corpora. Third, the number of unique syllables in each corpus varies considerably by language (e.g., Spanish: 522, Hungarian: 3,029). While some of this variation is due to the size of the corpus itself (more utterances allow more syllable types to appear), there are also language-specific phonotactic restrictions on syllables that impact the number of syllable types observed. For example, in Japanese only the phoneme /ɴ/ may appear after a vowel, and in Spanish only the phoneme /s/ can appear as the second consonant in a coda. In contrast, languages such as English, German, and Hungarian allow much more complex syllable types (e.g., consider the English coda in *warmth*, /ɹmθ/).

The average number of syllables per utterance also varies, which is partially due to speech directed at older children containing longer utterances (e.g., Farsi utterances have 6.98 syllables per utterance and are directed at children up to age five). However, the number of syllables per utterance is also impacted by the number of syllables per word – languages that tend to be more monosyllabic will tend to have fewer syllables per word, and so their utterances will tend to have fewer syllables as well. Boundary probability captures this monosyllabic tendency, where higher probabilities indicate that syllables tend to be followed by boundaries (i.e., words are more likely to be monosyllabic). For example, the English and German data have higher boundary probabilities than the other languages, and therefore tend to have more monosyllabic words.

8.4.2 Gold Standard Evaluation

8.4.2.1 Evaluation Metrics We first present the DP segmentation's ability to match the gold standard, the adult-level knowledge typically represented via the orthographic segmentation. There are multiple units a segmentation can be measured on: word tokens, lexical items, and word boundaries. For example, the utterance *The penguin is next to the kitty* might be segmented as *Thepenguin is nextto the kitty*. There are seven word tokens (individual words) in the original sentence, but the segmentation only identifies three of those tokens (*is*, *the*, and *kitty*). The same utterance has six lexical items, which correspond to the unique words: *the, penguin, is, next, to, kitty* (*the* appears twice). Again, the segmentation only correctly identifies three lexical items (*is*, *the*, and *kitty*). This utterance also has six word boundaries (excluding the utterance boundaries) and the segmentation correctly identifies four of those (*thepenguin is*, *is nextto*, *nextto the*, and *the kitty*). This highlights the differences between the units.

Word tokens are impacted by word frequency, while lexical items factor out word frequency. Measuring boundary identification tends to yield better performance than measuring word tokens or lexical items. Intuitively, this is because identifying a boundary requires the model to be correct only once. On contrast,

correctly segmenting words requires being correct twice, both in inserting the word-initial and word-final boundaries (unless the word is at an utterance edge).

No matter which unit we use for comparison, they are measured with the same metrics: precision and recall, which are often combined into a single summary statistic, the F-score. Precision captures how accurate the segmentation is: for each unit identified, is that unit correct in the segmentation? Recall captures how complete the segmentation is: for each unit that should have been identified, is that unit identified in the segmentation? In signal detection theoretic terms, these correspond to (8.10), which involve *Hits* (units correctly identified in the segmentation), *False Alarms* (units identified in the segmentation that aren't correct), and *Misses* (units not identified in the segmentation that are nonetheless correct).

$$Precision = \frac{Hits}{Hits + False\ Alarms} \qquad Recall = \frac{Hits}{Hits + Misses} \tag{8.10}$$

High precision indicates that when a unit is identified in a segmentation, it is often correct. High recall indicates that when a unit should be identified in a segmentation, it often is. Because both of these are desirable properties, they are often combined into the F-score via the harmonic mean. Precision, recall, and F-scores all range between 0 and 1 (though sometimes this is represented as a percentage between 0 and 100), with higher values indicating a better match to the gold standard.

$$F\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8.11}$$

8.4.2.2 Model Training and Parameter Estimation Because the algorithms used for inference are probabilistic in nature, each modeled strategy (DP-Uni and DP-Bi) was trained and evaluated five times, with the results averaged. Although the learning process modeled is unsupervised, Phillips and Pearl (2014a, 2014b) nonetheless separated the data into training and test sets to better determine how each modeled strategy adapted to new data. Each corpus was randomly split five times so that the training set consisted of 90% of the corpus and the remaining 10% became the corresponding test set. The corpora themselves are temporally ordered, so utterance order captures the order an infant might encounter the utterances. This relative ordering was preserved in both the training and test sets.

The free parameters for the DP-Uni (α) and DP-Bi strategies (β, γ) were set by searching ranges derived from those explored in Goldwater et al. (2009) and Pearl et al. (2011): α, $\beta \in [1, 500]$, $\gamma \in [1, 3000]$. For each language and each strategy, a learner using the idealized inference algorithm was used to

Table 8.2 *Best free parameter values for all unigram and bigram Bayesian segmentation strategies across each language.*

	DP-Uni	DP-Bi	
	α	β	γ
English	1	1	90
Italian	1	1	90
German	1	1	100
Spanish	1	200	50
Japanese	1	300	100
Farsi	1	200	500
Hungarian	1	300	500

determine which free parameter values resulted in the best word token F-score. The values identified by this process are shown in Table 8.2.

When we look at the best parameter values, it turns out that the DP-Uni strategy fares best on all languages when $\alpha = 1$. This indicates a strong bias for small proto-lexicons, since novel words are dispreferred. In contrast, the DP-Bi strategy has more variation, roughly breaking into three classes. The first class, represented by English, Italian, and German, has $\beta = 1$ and γ between 90 and 100. This indicates a very strong bias for small proto-lexicons, since novel bigrams are strongly dispreferred (β) and novel words as the second word in a bigram are somewhat dispreferred (γ). The second class, represented by Spanish and Japanese, has β between 200 and 300 and γ between 50 and 100. This indicates a weaker bias for small proto-lexicons, since novel bigrams are only somewhat dispreferred (β) and novel words as the second word in a bigram are also only somewhat dispreferred (γ). The third class, represented by Farsi and Hungarian, has β between 200 and 300 and $\gamma = 500$. This indicates an even weaker bias for small proto-lexicons, since novel bigrams are again only somewhat dispreferred (β) and novel words as the second word in a bigram are even less dispreferred (γ).

Since we are setting these parameter values for our analyses, this translates to the modeled infant already knowing the appropriate values for each language. For the DP-Uni strategy, this may reflect a language-independent bias, since the values are all the same. However, for the DP-Bi strategy, the infant would need to adjust the values based on the ambient language. For either strategy, one potential way to converge on the best values is to have hyperparameters on them and simply learn their values at the same time as segmentation is learned. Incorporating hyperparameter inference into the DP segmentation strategy would be

Table 8.3 *Word token F-scores for learners across English, German, Spanish, Italian, Farsi, Hungarian, and Japanese. Higher token F-scores indicate better performance, with the best score for each language in bold.*

		Eng	Ger	Spa	Ita	Far	Hun	Jpn
DP-Uni	Idealized	53.1	60.3	55.0	61.9	66.6	59.9	63.2
	Constrained	55.1	60.3	56.1	58.6	59.6	54.5	63.7
DP-Bi	Idealized	77.1	73.1	**64.8**	**71.3**	**69.6**	**66.2**	**66.5**
	Constrained	**86.3**	**82.6**	60.2	60.9	62.5	59.5	63.3
Baseline	RandOracle	56.4	47.5	27.0	22.8	20.3	26.4	26.1

a welcome avenue for future segmentation modeling work, particularly if it turns out infants are using something like the DP-Bi strategy. Here we discuss the results obtained by manually setting the free parameters to their respective optimized values in each language.

8.4.2.3 Baseline Strategy: Random Oracle A baseline strategy first explored by Lignos (2012) is a random oracle strategy (RandOracle). The random aspect refers to the strategy treating each possible boundary location as a Bernoulli trial. The oracle aspect comes from the strategy already knowing the true probability of a boundary occurring in the corpus (B Prob in Table 8.1). Boundaries are then randomly inserted with this probability.

8.4.2.4 Cross-Linguistic Performance Table 8.3 presents the gold standard word token F-score results for each learner on all seven languages. First, we can see that bigram assumption is generally helpful, though the degree of helpfulness varies cross-linguistically. For example, it seems very helpful in English and German (e.g., English Idealized DP-Uni: 53.1 vs. DP-Bi: 77.1; German Constrained DP-Uni: 60.3 vs. DP-Bi: 82.6) and not particularly helpful at all in Japanese (Japanese Idealized DP-Uni: 63.2 vs. DP-Bi: 66.5; Japanese Constrained DP-Uni: 63.7 vs. DP-Bi: 63.3). Still, with the exception of the English DP-Uni learner, every single Bayesian learner does better than the random oracle baseline. Interestingly, in English, the DP-Bi constrained learner has the highest score of all learners in all languages. Altogether, this suggests that the DP segmentation strategy is generally a very good one for identifying words in fluent speech.

Nonetheless, something seems to be going on cross-linguistically. Why do we see such variability in segmentation performance (DP-Uni: 53.1–66.6; DP-Bi: 59.5–86.3; RandOracle: 20.3–56.4)? The variability in the random oracle baseline is particularly suggestive that there is something about the languages

themselves, rather than something specific to the DP segmentation strategy. More specifically, English and German seem inherently easier to segment than the other languages.

Fourtassi et al. (2013) suggested that some languages are inherently more ambiguous with respect to segmentation than others. Specifically, even if all the words of the language are already known, some utterances can *still* be segmented in multiple ways (e.g., /gɹejtfʌl/ segmented as *great full* and *grateful* in English). The degree to which this occurs varies by language, with the idea that languages with high inherent ambiguity are harder to correctly segment. If this is true, we might expect that low inherent segmentation ambiguity correlates to high performance by segmentation strategies. With this in mind, perhaps English and German have lower inherent segmentation ambiguity than the other languages (RandOracle English: 56.4, German: 47.5, Other languages: 20.3–26.4).

In order to quantify this ambiguity, Fourtassi et al. (2013) proposed the normalized-segmentation entropy (NSE) metric:

$$NSE = -\sum_{s} P_s \log_2(P_s)/(N-1) \tag{8.12}$$

Here, P_s represents the probability of a possible segmentation s of an utterance and N represents the length of that utterance in terms of potential word boundaries (which is determined by the number of syllables for our learners). To calculate the probability of an utterance, we use the unigram or bigram DP generative model equations described in Section 8.3, since these represent the probability of generating that utterance under a unigram or bigram assumption. As an example, to calculate the NSE of a single utterance /aimgɹejtfʌl/, we use the unigram and bigram model equations to generate the probability of every segmentation comprised of true English words (P_s above). In this case, two segmentations are possible: *I'm grateful* and *I'm great full*. The probabilities for each segmentation are then used in Equation 8.12 above, with $N = 2$ since there are two potential word boundaries among the three syllables.

Because a low NSE represents a true segmentation that is less inherently ambiguous for the learners using the n-gram assumptions tested here, English and German should have lower NSE scores if inherent segmentation ambiguity was the explanation for the better segmentation performance. Table 8.4 shows the NSE scores for both unigram and bigram learners for all seven languages, with token F-scores for the respective idealized inference learners for comparison.

From Table 8.4, we see that German fits with the hypothesis that low NSE predicts higher segmentation performance, having in both cases the lowest NSE scores. Yet English does not fit this pattern, ranking third/fourth overall for a unigram DP learner and fourth overall for a bigram DP learner. This is despite

Table 8.4 *Average NSE scores across all utterances in a language's corpus, ordered from lowest to highest NSE and compared against the idealized inference token F-score for the language. Results are shown for both the DP-Uni and DP-Bi models. Lower NSE scores represent less inherent segmentation ambiguity and higher token F-scores indicate a better segmentation performance.*

DP-Uni	NSE	F-score	DP-Bi	NSE	F-score
German	0.000257	60.3	German	0.000502	73.0
Italian	0.000348	61.9	Italian	0.000604	71.3
Hungarian	0.000424	59.9	Hungarian	0.000694	66.2
English	0.000424	53.1	English	0.000907	77.1
Farsi	0.000602	66.6	Spanish	0.00103	64.8
Japanese	0.00126	55.0	Farsi	0.00111	69.6
Spanish	0.00128	63.2	Japanese	0.00239	66.5

English having the lowest token F-scores for the DP-Uni learner and highest token F-scores for the DP-Bi learner. Because of this, the high segmentation performance on both German and English cannot simply be due to both having lower inherent segmentation ambiguity.

More generally, it becomes clear by looking at all seven languages that low NSE does not always lead to higher token F-scores. If it did, we would expect to find a significant negative correlation between NSE score and token F-score – but this does not happen (DP-Uni: $r = -0.084$, $p = 0.86$; DP-Bi, $r = -0.341$, $p = 0.45$). Examining individual languages in Table 8.4, this lack of correlation is apparent. The DP-Uni Farsi NSE score is ranked fifth lowest, but in fact has the highest F-score, while the DP-Uni Spanish NSE score is actually the worst, though it has the second best F-score. When we turn to the DP-Bi learners, we see that Hungarian has the third best NSE score but the next to worst F-score, while English has the fourth worst NSE score but the best F-score. So, NSE is not the principal factor determining segmentation performance, though it may play some role.

An alternative factor comes from considering how often words of the language tend to be monosyllabic. This is captured by the boundary probability in Table 8.1, where English and German both have a much higher probability of having a boundary appear after a syllable (76.26% and 68.60%, respectively, compared to the next highest language Spanish, with boundary probability 53.93%). These are precisely the languages that especially benefit – though only for the DP-Bi and RandOracle learners.

One possible explanation for why boundary probability especially impacts these learners relates to the types of errors these learners make. More