Discovering syntactic hierarchies

Virginia Savova, Daniel Roy, Lauren Schmidt & Joshua B. Tenenbaum
{savova, droy, lschmidt, jbt}@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Abstract

The acquisition of syntactic rules is predicated upon the successful discovery of syntactic categories (parts of speech). These do not simply constitute a set, but in fact form a nested hierarchy, which allows rules to apply at different levels of generality. Languages provide a variety of cues to syntactic categorization - phonological, semantic and distributional. However, the relative reliability of these cues differs from language to language. This paper presents a computational model capable of acquiring the hierarchy syntactic categories from different combinations of cues. Interestingly, the model is domain general and has been successfully applied to non-linguistic discovery of hierarchical structure. **Keywords:** computational modeling; hierarchical clustering; linguistics; syntactic categories; language acquisition.

Introduction

Knowledge of syntax is knowledge of the combinatorial properties of words. Since it is not only infeasible, but outright impossible to encounter all licit combinations for any individual word, generalization over abstract categories is a crucial step in language acquisition. However, the task of uncovering the categorial structure of lexical items is highly non-trivial. While members of the same category share certain semantic or morpho-phonological characteristics, there is no guarantee that items with shared characteristics fall in the same category. For example, all count nouns take the suffix 's' (to denote plural), but so do all verbs (to denote 3rd person singular). Similarly, members of the same syntactic category may differ widely in both meaning and sound (e.g. 'salt' and 'furniture'). It follows that a syntactic category is best defined by abstract combinatorial properties. This leads us to a classical chicken-and-egg problem: while the acquisition of syntactic rules is predicated upon the successful discovery of syntactic categories, the categories are in turn identified on the basis of these rules. How could human learners extricate themselves from this predicament?

Before we present our approach, let us take a closer look at the nature of syntactic categorization. While many researchers make the simplifying assumption that the structure of categories is flat (e.g. (Cartwright & M., 1997), (Clark, 2003)), it is better to conceive of them as organized in a nested hierarchy. This organization allows combinatorial rules to be associated with different levels of generality within the hierarchy. For example, all English verbs require a subject, but only a subset of verbs require an object (the so-called transitive verbs, e.g.'hit'). Similarly, while all nouns share some combinatorial properties, only common nouns (e.g. 'salt', 'book', but not 'John') can occur with a definite determiner ('the'), and only a subset of these (e.g. 'book', but not 'salt') can occur with an indefinite determiner ('a').

While participation in combinatorial rules is the defining characteristic of a syntactic category, it is often the case that members of the same category tend to share semantic properties. Verbs tend to refer to events, nouns – to objects or people. Both developmental psychologists and linguists have argued that semantic cues play a significant role in the early stages of syntactic development. In particular, Macnamara (1972) proposed that children acquire syntactic knowledge on the basis of already developed knowledge of concepts and semantic relations. Later, Pinker (1982) suggested that children use their understanding of verb meaning to infer the syntactic frames in which they appear and vice versa. This is probably facilitated by the consistency of caregiver speech with respect to semantic-syntactic mapping (Rondal & Cession, 1990). Cross-linguistic typologies of case (Grimshaw, 1981) can also be accounted for by postulating an innate mapping preference from agents to subjects of active sentences.

In addition, a category may be marked by overt morpho-phonological markers. The reliability of this type of cue varies greatly from language to language. For example, the English suffix 'tion' applies exclusively to the noun class, and overwhelmingly to abstract nominals (define-definition, prescribe-prescription etc.). However, English rarely marks the syntactic category of words, as the existence of identical noun-verb pairs attests (e.g. to chase – a chase, to jump – a jump etc.). In contrast, a morphologically rich language (e.g. Russian), provides a wide variety of suffix and inflectional cues that distinguish syntactic categories.

Ultimately, distributional information is paramount and the contributions of semantics and phonology must be reconciled with it. While previous approaches treat contextual cues as a type of lexical feature, cooccurence is best described as a binary relation. This intuition is captured by virtually all grammar formalisms, including dependency grammar, varieties of phrase-structure grammar, LFG, X-bar theory and minimalism. In fact, many formalisms postulate more than binary relation among words. Thus, acquiring categories from distributional cues is a special case of identifying categories from multiple relational cues.

All of this suggests that learning the hierarchical structure of syntactic categories is a complex process involving the integration of many cues, which fall into two major classes: relational and feature-based. Feature-based cues involve the presence of semantic or morpho-phonological information associated with lexical entries. Relational cues involve the membership of lexical pairs in certain types of (distributional) relations. Since different languages employ feature-based cues to different extent, it is important for a computational

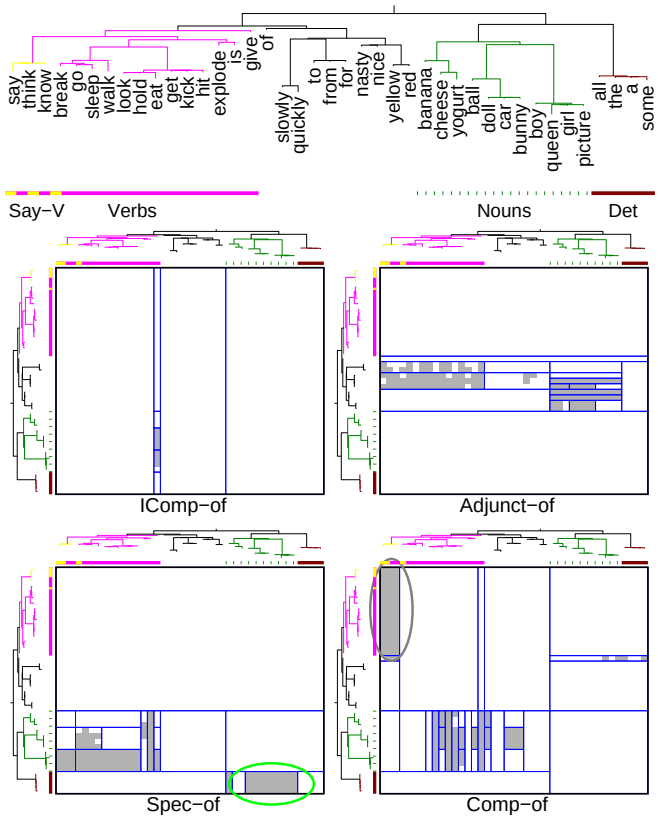

Figure 1: Tree and relation matrix for the X-bar dataset: Spec-of relation matrix illustrates the association between determiners and nouns on one hand (circled in green), and verbs and nouns on the other. The Comp-of relation matrix associates a subclass of verbs—the reflective verbs (Say-V) with the verb class as a whole (circled gray).

model to be general enough to profit from these cues if and when they are available, while being able to deal with the absence of these cues when unavailable. Thus, a model should be able to naturally incorporate multiple sets of both relational and feature-based data. Our method for discovering annotated hierarchies (Roy, Kemp, Mansinghka, & Tenenbaum, 2006) was developed specifically for learning situations of this sort. This is the first application of the model to linguistic data.

Annotated hierarchies model

Given a collection of word features (e.g. morphological or semantic), an *annotated* hierarchy specifies nested categories of words, as well as the appropriate categories with which to summarize the observed features. For example, an annotation for a feature indicating whether a word can take the "-ing" suffix would specify that the category containing all verbs has this property while the three categories containing all nouns, all adjectives and all adverbs do not. Many syntactic properties are best described by relations between words and annotation hierarchies summarize the observed relations by specifying how certain categories of words relate to one another. For example, consider the relations *object-of* and *subject-of*; while all verbs require a subject, only transitive verbs require an object. Therefore, the appropropriate level in the hierarchy to describe the *subject-of* relation is the category of all verbs but the a finer-grained distinction of transitive/intransitive is relevant for the relation *object-of*.

The idea of an annotated hierarchy is one of the oldest proposals in cognitive science, and researchers including Collins and Quillian (1969) and Keil (1979) have argued that semantic knowledge is organized into representations of this form. Previous treatments of annotated hierarchies, however, often suffer from two limitations. First, annotated hierarchies are usually hand-engineered, and there are few proposals describing how they might be learned from data. Second, annotated hierarchies typically capture knowledge only about the features of objects: relations between objects are rarely considered. In contrast, our generative probabilistic model simultaneously handles objects, features, relations, and can be used to recover annotated hierarchies from raw data.

The annotated hierarchies model assumes that the objects are located at the leaves of a rooted tree (each node specifies the category of objects in its subtree), and that each feature and relation is generated independently conditioned on the structure of the tree. Intuitively, objects that are nearby in the tree will tend to have similar features values, and relate to other objects in similar ways. In this setting, objects are words and we are trying to discover an annotated hierarchy of these words that summarizes the observed morphological features and syntactic relations. More precisely, each feature (or relation) is associated with a partition of all words (or of all pairs of words) and this partition is constrained to respect the tree structure (i.e. each subset in the partition is an entire category specified by the hierarchy). Therefore, one can think of these partitions as lists of categories (or pairs of categories in the relational case). The model contains a prior over partitions that encourages partitions to use the most general categories possible without losing too much predictive accuracy.

Each category (or pair of categories) in a partition is associated with a real-valued parameter θ between 0 and 1 that specifies the probability with which the feature (or relation) applies to words (or pairs of words) in that subset. These "parameterized" partitions describe the typical values for each feature and relation for different branches of the tree. For example, the category of all verbs would likely be included in the partition describing the "-ing"-suffix feature and the corresponding parameter would be closer to 1 than 0 because many verbs would be observed as gerunds at some point. A parameterized partition associated with a relation describes how likely it is that any pair of words stand in that relation, as a function of the location of the words in the hierarchy.

The probability of the i'th feature, F_i, conditioned on the tree T, can be computed by summing the contribution of every possible partition π, weighted by its prior probability: $P(F_i|T) = \sum_\pi P(F_i|\pi)P(\pi|T)$. In the same manner, we can

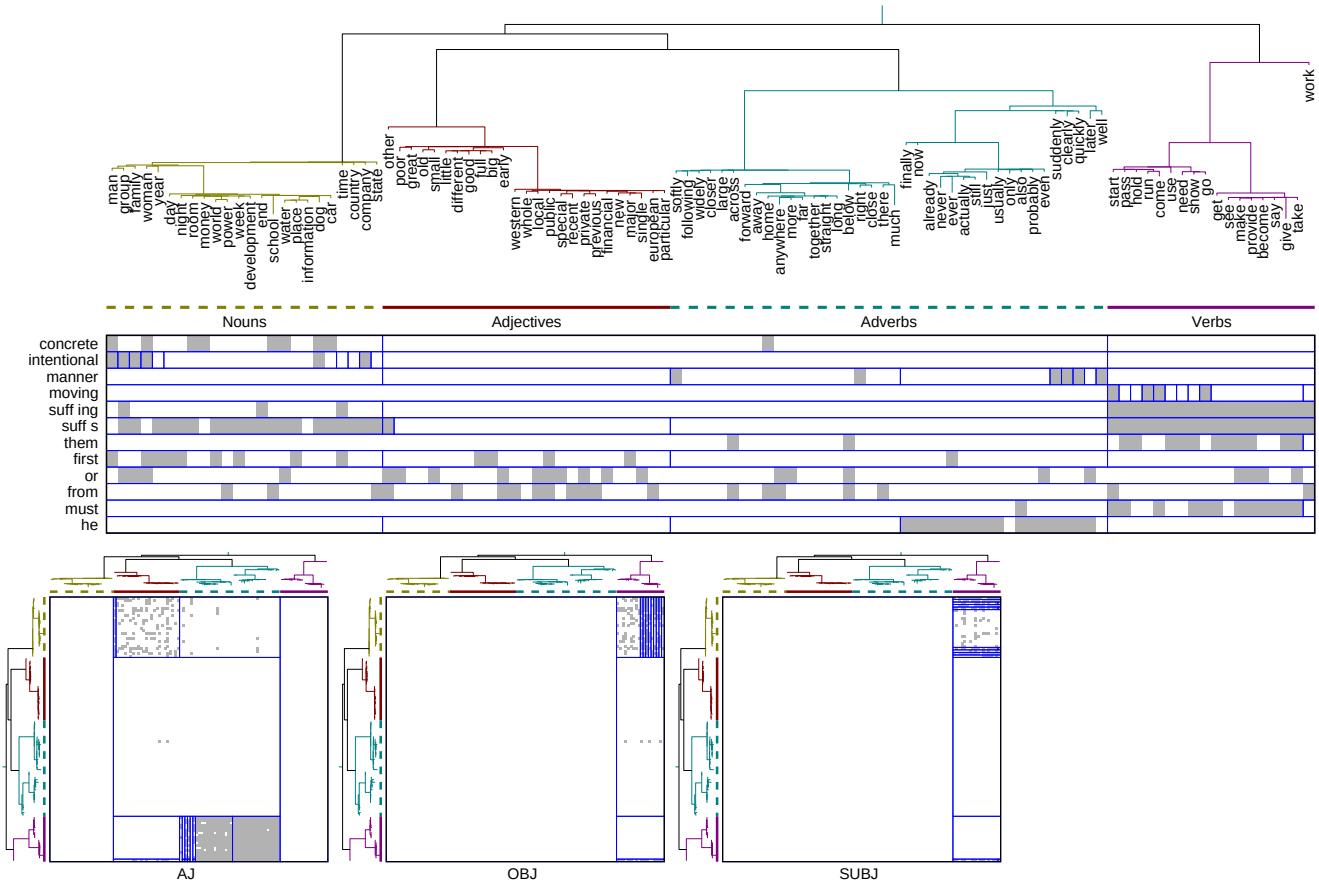

Figure 2: Trees induced from BNC data: a) relations, local context features, semantic and morphological features.

compute the probability of the j'th relation, R_j, conditioned on the tree T: $P(R_j|T)$. Given features $F_1,\ldots,F_n$ and relations $R_1,\ldots,R_m$, the posterior probability of a tree T is given by Bayes' rule as

$$P(T|F_1,\ldots,F_n,R_1,\ldots,R_m) \propto P(T)\prod_i P(F_i|T)\prod_j P(R_j|T),$$

where $P(T)$ is a prior over tree structures. Roughly speaking, the best hierarchy will then be the one that provides the best categories with which to summarize all the features and relations. For lack of space, we refer the reader to a recent publication where the model is presented in full detail (Roy et al., 2006).

Related work

Induction of syntactic categories

Part-of-speech tagging is a highly successful application of statistical Natural Language Processing (NLP) techniques. However, our work differs from NLP research in fundamental ways. First, the goal of PoS tagging is to label text with the best tag of a pre-defined set, rather than inducing the categories themselves. Second, since the goal is to create an accurate engineering application against a particular benchmark, learning is always supervised. Third, no hierarchical structure of tags is assumed, and the pre-defined set of labels can be viewed as a low-level horizontal cut through the actual hierarchy. While the literature on PoS tagging is largely orthogonal to our approach, other attempts of unsupervised clustering on NLP data provide useful comparisons. In particular, clustering algorithms have been applied to induce semantic categories in tasks such as word sense disambiguation, and identification of word senses. Dekang Lin's work (Lin, 1998) is particularly interesting in this regard. Using relational data for multiple dependency relations, he is able to identify pairs of semantically related words. Although the intention is to obtain words with similar meanings, the resulting pairs are also close syntactic neighbors. However, there is no notion of hierarchy and association with particular relations.

Unsupervised induction of PoS categories in NLP is relevant in the context of cross-linguistic applicability, which necessitates combining morphological and distributional information. The most important difference with respect to our proposal is the resulting category structure. The goal of NLP approaches is to induce a flat set of categories, rather than a categorial hierarchy. In addition, the number of clusters is set in advance. Clark (2003) presents a series of cross-linguistic experiments in unsupervised PoS induction with Hidden Markov Models, based on a combination of distri-

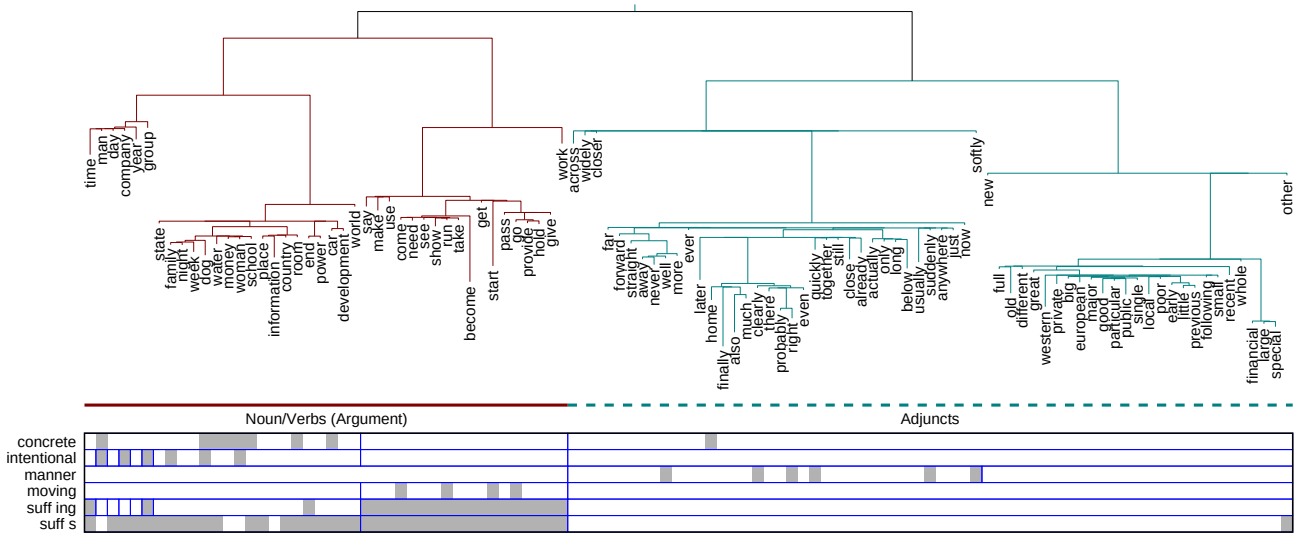

Figure 3: Trees induced from BNC data: single dependency relation + semantic and morphological features.

butional and morphological information. He demonstrates that the inclusion of morphological information improves the clustering of rare words in morphologically rich languages. There is no natural way of modifying his approach to result in hierarchical structure.

The role of distributional information in syntactic categorization has received a reasonable amount of attention in cognitive science. In some cases a categorial hierarchy is induced. The work of Redington, Chater, and Finch (1998), is among the most detailed in this respect. By applying hierarchical clustering methods on the distributional contexts of words from the CHILDES corpus of caregiver speech, they produce a dendrogram which captures the main part-of-speech classes (noun, verb, adjective), along with some substructure. Similar results have been obtained by others with the same approach for nouns and verbs but in a smaller distributional context (Mintz, Newport, & Bever, 2002). A slightly different variation was developed by Jeffrey Elman on a corpus of short sentences generated by a simple artificial grammar. He trained a simple recurrent neural network to predict the next word of the input. The units of the network were treated as a feature vector a subsequent hierarchical cluster analysis (Elman, 1991), which showed some representation of the underlying word classes. Another approach to distributional clustering was investigated by Cartwright and Brent (Cartwright & M., 1997). While their method has certain advantages over hierarchical clustering (e.g. it works incrementally), its main drawback is that it results in a discrete set of categories and does not capture the nested structure of categories. Unlike our method, hierarchical clustering relies on only one source of information at a time and cannot combine knowledge of multiple relations and/or features to produce the best representation for all. Furthermore, this type of clustering results in an enormous number of nested subcategories, most of which have no natural interpretation.

Previous applications of annotated hierarchies model

In previous work, the model was successfully applied to non-linguistic cognitive tasks (Roy et al., 2006). It was shown to discover the conceptual structure of feature data from four domains: animals, food, vehicles and tools. In addition to identifying the four domains, the model came up with relevant superordinate and subordinate categories. It was also successful in uncovering the kinship structure of Australian tribes.

Experiments

In the first experiment, a small number of simple sentences were used to extract the three basic cooccurence relations postulated by the X-bar theory of syntax. According to X-bar theory, each word may select $1-2$ obligatory arguments (*specifier* and *complement*), and an unbounded number of optional *adjuncts*. The traditional substantive relational categories of verb and object are interpreted as special cases of the specifier and complement relations, which are not limited to verbs alone. For example, most English common nouns require a determiner in the specifier position, just as verbs require a noun in the subject position. Analogously, some nouns require prepositional phrases as complements, just as verbs require objects (e.g. the noun "picture" requires a complement "of X" to be interpreted). In addition, a small subset of verbs (e.g. give) require a secondary obligatory argument (IComp), which refers to a beneficiary or recipient. If language learners are able to observe these fundamental relationships at the word-to-word level, would they be able to use the relational data to form a hierarchical structure of categories?

To answer this question, we picked forty words that are likely to figure into early vocabulary, and represent an interesting set of potential subcategories. These included nouns, verbs, adjectives, adverbs, prepositions and determiners. The

results of our experiment show that the algorithm uncovers linguistically relelvant structure at multiple levels (Figure 1). It splits the words into nouns, verbs, adjectives, determiners, prepositions and adverbs and identifies important subclasses within the verb class: a subclass of reflective verbs ('think', 'say', 'know') and verbs that are used without direct object (walk, sleep, break, go), as well as a set of transitive verbs (hit, kick, hold, get and eat). It places the ditransitive 'give' in a separate category. Subclasses are also identified in the noun category. It is roughly split into physical objects on one side, and people on the other. An exception is 'picture', which clusters with the people category (cf. 'queen of'/'picture of').

Corpus-based experiments

The success of the first experiment lead us to explore more realistic scenarios using automatically extracted data from a real corpus, and less abstract relations which are easier to observe on the surface. Ultimately, the relations postulated by X-bar theory are generalized versions of subject/object/modifier relations which have a semantic basis. In addition, these relations have a strong reflection on surface order—in English, subjects almost always precede—and object follow, the verb. Our next set of experiments is with a dataset of automatically extracted subject, object and modifier relations for categorially heterogeneous set of words. To obtain a dataset of manageable proportions, the relations were collected with a frequency threshold. The set of words was chosen on the basis of frequency with two criteria: First, to contain an approximately equal number of items from the four main categories (Nouns, Verbs, Adjectives, and Adverbs), and second, to create a reasonably dense relation matrix, so that each word relates to at least three other items. The resulting set contained 105 words: 18 nouns, 29 adjectives, 36 adverbs, and 22 verbs.

We collected the data from the British National Corpus (BNC), a corpus of 100M words. The Phrases In English (PIE) utility (available at `http://pie.usna.edu`) allowed us to find the most popular nouns in the BNC by using its n-gram tool to identify the most popular unigrams containing common nouns. We selected a subset of words from this set and added a few more common words to better span the ontological categories of concrete nouns and living things. Using another online utility, the Sketch Engine (Kilgarriff, Rychly, Smrz, & Tugwell, 2004, available at `http://www.sketchengine.co.uk/`), we produced word sketches of these nouns – summaries of their relationships to other words within the BNC and the frequencies with which they occur in these grammatical and collocational relations. We collected data on which words pairs occurred in "object of" relations (e.g., $object\text{-}of(money, give)$), "subject of" relations (e.g., $subject\text{-}of(I, give)$), and modifier relations (e.g., adjective modifiers like $modifies(tall, man)$). From this data, we selected a subset of verbs and adjectives that were densely relationally connected to the original nouns, and we repeated the word sketch process on these words, additionally pulling out adverbial modifier relations (e.g., $modifies(suddenly, go)$), which were collapsed with the adjectival modifer relations into a single relation for all of the corpus-based data sets.

One of the interesting questions we set out to investigate whether the algorithm can discover internal structure in the relations, when the structure is not explicitely given. To do so, we collapsed the two types of modifier relations provided by the relation extraction utility into a single relation. We also ran an experiment where all relations were collapsed into one dependency relation ('a cooccurs with b', where a and b are arbitrarily far from one another), and an experiment based on immediate adjacency ('a adjacent to b').

We also investigated the effect of manually annotated semantic and morphological features. Semantic features were shared by a small subset of items in the major categories. One morphological feature ('ing') was highly consistent with the verb category, but also occured with items in the noun category, which happen to have dual status ('work', 'time'). The second morphological feature however ('s') was evenly split between the noun and the verb category, denoting plural inflection with nouns and 3rd person singular with verbs. The question is whether the algorithm would be able to assign the correct level of importance to these features, overriding their effect when distributional relations strongly favor other partitions. In addition, we ran experiments using bigram distributional features akin to those typically used in hierarchical clustering, and a set of experiments with the standard hierarchical clustering algorithm on the same data. The PIE utility allowed us to find the most common bigrams containing our words of interest. We gathered cooccurrence data using this tool and represented the data as features of the form *preceded-by-X* and $followed\text{-}by\text{-}Y$. A subset of the datasets was then selected so as to contain a densely connected set of popular words.

Results and discussion

In all of our experiments, the algorithm was successful in identifying the high level syntactic categories. This is particularly impressive when all relations identified were collapsed into one. Adjectives and adverbs were always recovered as distinct categories, regardless of the fact that both participate in the modifier relation (Figure 2). In addition, when all relations were collapsed into one, the algorithm found high level structure, separating the adjective/adverb superclass from the noun/verb superclass (Figure 3). This is interesting given the linguistic relevance of distinguishing optional and obligatory elements. With the exception of the first experiment however, the algorithm found relatively little subordinate structure. Nevertheless, the success of the first experiment (Figure 1) convinces us that the reluctance to separate subgroups is due to noise in the corpus data.

In general, the model made better use of semantic features in identifying subgroups when fewer relations were present. For example, the motion verbs cluster in a subgroup when all relations are collapsed into one, but fail to conclusively

separate when multiple relations are present. This is because the model places relatively little value on features versus relations. However, in reality subgroups are likely to share more than one feature, and including more of these features will probably improve performance. The phonological features introduced in the model help classification in so far as they do not contradict relational evidence. The 's' suffix feature is linked to two clusters, noun and verb, the 'ing' feature – to the verb cluster.

The comparison with hierarchical clustering (HC) using local context produced comparable results. While HC is slightly better at picking out low-level subclasses, it has no principled way of associating levels of the hierarchy with relations. Therefore, the results of HC have less predictive power than the representations derived by our model. In fact, one might argue that the difference in performance is due to issues of implementation. Improving the algorithm's ability to handle noise will allow for a better comparison on corpus data. This is supported by a comparison with HC on the original dataset of X-bar relations, where our algorithm performs slightly better.[1]

It is worth pointing out that the annotated hierarchies model was originally developed to handle data from other cognitive domains. Thus, it can be considered a domain-general mechanism for acquiring hierarchical structure for the purposes of probabilistic reasoning. It is remarkable that the problem of learning syntactic categories appears to share high-level similarities with concept development.

Future work

We are currently working on a number of additional experiments intended to investigate the performance of the algorithm on relations and features automatically extractable from raw data. In particular, we would like to replace our results on subject-of and object-of relations with the ordering relations 'precedes' and 'follows'. One issue is that it is not feasible to use linear order naively, since many important ordering relations are not adjacent. For example, the order of the verb and its object is often interrupted by a determiner (as in 'ate an apple'). While this is a problem for a naive automatic text analysis, there are reasons to believe it does not present such a problem for children, who are probably able to filter out low-saliency unstressed words from familiar content words. Thus, extracting precedence relations involves finding an appropriate way of automatically filtering important words. Another direction we are pursuing is the automatic identification of semantic (pragmatic) and morphological features. Initially, the latter can be accomplished in a supervised way, while the ultimate goal is to rely on completely unsupervised extraction. Our strategy with respect to the semantic features is to obtain a corpus annotated with pragmatic cues (agent, action, patient, goal). Last but not least, we are working on extending our model to scale up to larger datasets in order to improve the hierarchies at the subordinate level, and to provide a more extensive comparison to hierarchical clustering methods. This involves increasing the model's tolerance to noise, and developing faster inference methods.

[1] For example, HC clusters 'of' with the verbs.

References

Cartwright, T., & M., B. (1997). Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition*, *63*(2), 121-170.

Clark, A. (2003). Combining distributional and morphological information for part of speech induction. In *Eacl'03: Proceedings of the tenth conference on european chapter of the association for computational linguistics* (pp. 59–66).

Collins, A. M., & Quillian, M. R. (1969). Retrieval Time from Semantic Memory. *JVLVB*, *8*, 240–248.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195-224.

Grimshaw, J. (1981). Form, function, and the language acquisition device. In C. Baker & J. McCarthy (Eds.), *The logical problem of language acquisition* (p. 165-182). MIT Press.

Keil, F. C. (1979). *Semantic and conceptual development*. Cambridge, MA: Harvard University Press.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The sketch engine. In *Proceedings of euralex.* Lorient.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on computational linguistics.*

Macnamara, J. (1972). *Cognitive basis of language learning in infants.* (Vol. 79).

Mintz, T., Newport, E., & Bever, T. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, *26*, 393-424.

Pinker, S. (1982). A theory of the acquisition of lexical interpre-tive grammars. In J. Bresnan (Ed.), *The mental representation of grammatical relations* (p. 655-726). MIT Press.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information:a powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 425-469.

Rondal, J. A., & Cession, A. (1990). Input evidence regarding the semantic bootstrapping hypothesis. *Journal of Child Language*, *17*, 711-717.

Roy, D., Kemp, C., Mansinghka, V., & Tenenbaum, J. (2006). Learning annotated hierarchies from relational data. In *Proceedings of neural information processing systems 19 (NIPS).*