A graph-theoretic model of lexical syntactic acquisition

Hinrich Schütze and Michael Walsh
Institute for Natural Language Processing
University of Stuttgart, Germany
{hs999,walsh}@ifnlp.org

Abstract

This paper presents a graph-theoretic model of the acquisition of lexical syntactic representations. The representations the model learns are non-categorical or graded. We propose a new evaluation methodology of syntactic acquisition in the framework of exemplar theory. When applied to the CHILDES corpus, the evaluation shows that the model's graded syntactic representations perform better than previously proposed categorical representations.

1 Introduction

In recent years, exemplar theory has had great explanatory success in phonetics. Exemplar theory posits that linguistic production and perception are not mediated via abstract categories, but that instead each production and perception of a linguistic unit is stored and retained. Linguistic inference then directly operates on these stored *exemplars*. In this paper, we propose a new approach to lexical syntactic acquisition in the framework of exemplar theory.

Our approach uses an evaluation measure that is different from previous work. Lexical syntactic acquisition is most often evaluated with respect to standard syntactic categories like verb and noun. Our first contribution in this paper is that we instead evaluate learned representations in the context of a syntactic task. This task is the determination of an aspect of grammaticality that we call *local syntactic coherence*.

Our second contribution is a *graph-theoretic model of the acquisition of lexical syntactic representations* that is more rigorous than previous heuristic proposals. The graph-theoretic model can learn both categorical and non-categorical (or graded) representations. The model is also a unified framework for syntagmatic and paradigmatic relations (as will be discussed below), and for lower-

order syntactic relations (those that can be directly observed from the input) and higher-order syntactic relations (those that require some generalization from what is directly observable).

Redington et al. (1998) give an influential account of the acquisition of lexical syntactic representations in which a standard syntactic category like verb or noun is assigned to each word. Our third contribution is to show that, in the context of acquisition, *graded representations are superior to standard categorical representations* in supporting judgments of local syntactic coherence. A graded representation formalism is one that, for any two words, can represent a third word whose syntactic properties are intermediate between the two words (Manning and Schütze, 1999).

Clearly exemplar theory is not the only framework in which lexical acquisition has been explored. Gleitman (1990) for example argues for syntactic bootstrapping to infer lexical semantics, work not at odds with our own (see discussion on the role of semantics below). Our argument for the importance of distributional evidence does not call into question the large body of work in child language acquisition that demonstrates that "part of the capacity to learn languages must be 'innate'" (Gleitman and Newport, 1995). Tabula rasa learning is not possible. Our goal is not to show that language acquisition proceeds with a minimum of inductive bias. Rather, we attempt to formalize one aspect of language acquisition, the use of distributional information.

The paper is organized as follows. Section 2 motivates the exemplar-theoretic approach by reviewing its success in phonetics. Section 3 defines local syntactic coherence, which is the basis for a new evaluation methodology for the acquisition of lexical representations. Section 4 develops the graph-theoretic model. Section 5 compares graded and categorical representations for the task of inferring local syn-

tactic coherence. Section 6 presents our evaluation. Sections 7 and 8 discuss related and future work, and present our conclusions.

2 Exemplar theory

The general idea of research into exemplars in speech production and perception is that encountered items (segments, words, sentences etc.) are stored in great detail in memory along with rich linguistic and extra-linguistic context information. These exemplars are organized into clouds of memory traces with similar traces lying close to each other while dissimilar traces are more distant. A number of such models have had great success in accounting for production and perception phenomena in phonetics. E.g., Johnson (1997) offers an exemplar model which challenges the notion that speech is perceived through a process of normalization whereby a speaker-specific representation is mapped or normalized into a speaker-neutral categorical abstraction. Johnson's model successfully treats aspects of vowel perception, sex identification, and speaker variability. Crucially, no normalization of percepts into categorical representations takes place. The correct identification of phonemes and words in his model is a function of direct comparison to richly detailed exemplars stored in memory. Other examples of exemplar-theoretic phonetic accounts include (Goldinger, 1997), (Pierrehumbert, 2001), and our own work (Schütze et al., 2007). Exemplar theory's success in phonetics motivates us to investigate its use as a model for local syntactic phenomena.

3 Local syntactic coherence

In the context sequence model for exemplar-theoretic phonetics (Wade et al., 2008), we represent speech using amplitude envelopes derived from the acoustic signal and then compute similarity as the integral over the correlation of the two acoustic signals.

For the syntactic level, we need a representation that has two key properties of the representation we use in phonetics in order to support an exemplar-theoretic account. First, the representation must be directly derivable from the perceived input. In particular, it cannot rely on the results of any disambiguation that would occur either as part of exemplar-theoretic perception or in further downstream processing. Second, it must support similarity computations. Accordingly, we first motivate the representation we use and then introduce a similarity measure on these representations.

Representation. There are two main sources[1] of directly observable information about the syntactic properties of words: semantic cues (e.g., things are often referred to with nouns) and the neighbors of a word in sentences that it is used in. In this paper, we only consider the second source of information for acquisition, lexical neighbors.[2] We further limit ourselves to the immediate left and right lexical neighbors (see discussion in Section 7).

When using lexical neighbors as the basis of representation, we have to make a basic choice as to whether we look at left and right neighbors separately or whether we only look at the "correlated" neighborhood information of left and right neighbors jointly. Our approach is based on the first alternative: we separate the processing of left and right neighbors. We do this for two reasons. First, generalization improves and model complexity decreases if left-neighbor information and right-neighbor information are looked at separately. E.g., the right neighbors of *to*, *might* and *not* are similar because all three words can be followed by base verbs like *dance*: *to dance*, *might dance*, *(might) not dance*. But their left neighbors are very different.

Second, exemplar-theoretic similarity is best defined at the smallest possible scale in order to allow optimal matching between parts of the stimulus and parts of memory. In phonetics, we use a time scale of 10s of milliseconds or even less. Conceivably, one could also use segments (e.g., consonants and vowels) as the smallest unit; however, this would presume a segmented signal. And segmentation is part of the perception task we want to explain in the first place.

Separating left and right neighbors – which amounts to looking at left and right local contexts of each word separately – is the smallest scale we can operate at when doing syntactic matching. We

---

[1] A comprehensive account of acquisition must also include morphology. See Christiansen et al. (2004).

[2] Psycholinguistic evidence for the importance of neighbor information for learning categories includes (Mintz, 2002).

choose this small scale for the same reasons as we choose a small scale in phonetics: to ensure maximum flexibility when matching parts of the stimulus with exemplars in memory. Using words, bigrams or larger units would reduce the flexibility in matching and require a larger amount of experience (or training data) to learn a particular generalization.

We refer to the representations of left and right contexts of a given word as *half-words*. In other words, we split a word into two entities, a left half-word that characterizes its behavior to the left and a right half-word that characterizes its behavior to the right. Thus left-context and right-context components of the representation of a given focus word are defined, where a left (right) half-word consists of a probability distribution over all words that occur to the left (right) of the focus word and the dimensionality of the vector for each word is dependent on the number of distinct neighbors (left and right). For example, having experienced *take doll* twice and *drop doll* once, then the left context distribution, or left half-word of *doll*, $doll_l$, is $P(\text{take}) = 2/3, P(\text{drop}) = 1/3$. By extension, the phrase *take the doll* is represented as the following six half-words: $take_l$, $take_r$, the_l, the_r, $doll_l$, and $doll_r$.

Distance measure. The basic intuition behind local syntactic coherence is that an important component of syntactic wellformedness – and a component that is of particular importance in acquisition – is whether a similar sequence has already been stored as grammatical in memory. The same way that a phonetic signal that is well-formed in a particular language has many similar exemplars in memory, a syntactic sequence should also be licensed by similar, previously perceived sequences in memory. To operationalize this notion, we need to be able to compute the similarity or distance between an input stimulus and exemplars in memory. We do this by first defining a distance measure for sequences of fixed length.

The distance Δ between two sequences of half-words $< g_1, \ldots, g_n >$ and $< h_1, \ldots, h_n >$ is defined to be the sum of the distances of their half-words:

$$\Delta(<g_1, \ldots, g_n>, <h_1, \ldots, h_n>) = \sum_{i=1}^{n} \Delta(g_i, h_i)$$

This definition presupposes a definition of the distance of two half-words which will be given below.

We then call a sequence of n half-words $g_1, \ldots, g_n$ *locally coherent* if there is a sequence $h_1, \ldots, h_n$ in memory with $\Delta(< g_1, \ldots, g_n >, < h_1, \ldots, h_n >) < \theta$ where θ is a parameter.

Finally, we define a sentence to be *locally n-coherent* if all of its subsequences of length n are locally coherent.

The graph-theoretic model that is introduced in the next section will be evaluated with respect to how well it captures local syntactic coherence. This enables us to evaluate the model with respect to a task as opposed to its ability to reproduce a particular linguistic representation of syntactic categories.[3] Obviously, the notion of local syntactic coherence only captures some aspects of syntax – e.g., it does not capture long-distance dependencies. However, it is a plausible component of syntactic competence and a plausible intermediate step in the acquisition of syntax.

4 Graph-theoretic model

We briefly review the structuralist notions of syntagmatic and paradigmatic relationships that have been frequently used in prior work in NLP (e.g., (Church et al., 1994)). De Saussure defined a syntagmatic relationship between two words as their contiguous occurrence in a sentence and a paradigmatic relationship as mutual substitutability (de Saussure, 1962) (although he used the term *rapport associatif* instead of *paradigmatic*). E.g., *brown* and *dog* stand in a syntagmatic relationship with each other in the phrase *brown dog*; *brown* and *black* stand in a paradigmatic relationship with each other with respect to the position between *the* and *dog* in the phrase *the X dog*. De Saussure's conceptualization of syntactic relationships captures the fact that both admissible *neighbors* and admissible *substitutes* in language are an important part of the characterization of the syntactic properties of a word.

We formalize the two relations as *distributions over words*, where we assume a vocabulary $\{w_1, \ldots, w_V\}$ and V is the number of words in the vocabulary.

We denote the *left syntagmatic distribution* of w_i

---

by $p_{i,s,l,m}$ where i is the vocabulary index of w_i, s stands for *syntagmatic*, l for *left* and m is the order of the distribution as discussed below. Intuitively, $p_{i,s,l,m}(w_j)$ is the probability that word w_j occurs to the left of w_i. Similarly, for the *left paradigmatic distribution* of w_i, $p_{i,p,l,m}(w_j)$ is the probability that w_j can be substituted for w_i without changing local syntactic coherence as far as the context to the left is concerned. Note that we distinguish between left and right paradigmatic distributions. A word w_j can be a perfect substitute for w_i as far as the context to the left is concerned, but a very unlikely substitute as far as the context to the right is concerned. E.g., in the phrase *She loves her job*, the word *him* is a good left-context substitute for *her*, but a terrible right-context substitute for *her*.

We will now show how the syntagmatic/paradigmatic (henceforth: syn/para) distributions are defined iteratively, based on the bigram distribution p_{ww}, and grounded by defining $p_{i,p,l,1}$ and $p_{i,p,r,1}$.

$p_{ww}(w_i w_j)$ is the probability that the bigram $w_i w_j$ occurs, that is, that w_i and w_j occur next to each other (and in that order). We define the $V \times V$ *joint probability matrix* J by $J_{ij} = p_{ww}(w_i w_j)$.

Denote by N the diagonal $V \times V$ matrix that contains in N_{ii} the reciprocal of $p_w(w_i)$ where p_w is the marginal distribution of p_{ww}:

$$\sum_{j=1}^{V} p_{ww}(w_i w_j) = \sum_{j=1}^{V} p_{ww}(w_j w_i) = p_w(w_i) = \frac{1}{N_{ii}}$$

The conditional probability p_{left} of the following word and the conditional probability p_{right} of the preceding word can be computed by multiplying (the transpose of) J and N: $p_{\text{left}}(w_i|w_j) = p_{ww}(w_i w_j)/p_w(w_j) = (JN)_{ij}$; and $p_{\text{right}}(w_i|w_j) = (J^T N)_{ij}$.

The "grounding" paradigmatic distributions of order 1 are defined as follows.

$$p_{i,p,l,1}(w_j) = p_{i,p,r,1}(w_j) = \begin{cases} 0 & \text{if } w_i \neq w_j \\ 1 & \text{if } w_i = w_j \end{cases}$$

In other words, each word has only one perfect left / right substitute and that perfect substitute is itself. We define the syn/para distributions of higher order recursively:

$$p_{i,s,l,m} \quad = \quad JN p_{i,p,l,m} \qquad (1)$$

$$
\begin{array}{cc}
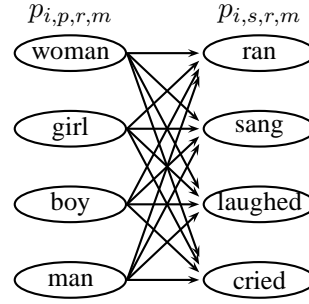p_{i,p,r,m} & p_{i,s,r,m}
\end{array}
$$

Figure 1: The distribution of typical right neighbors (the right syntagmatic distribution $p_{i,s,r,m}$) is computed from the distribution of typical "right substitutes" (the right paradigmatic distribution $p_{i,p,r,m}$).

$$p_{i,p,l,m} \quad = \quad J^T N p_{i,s,l,m-1} \qquad (2)$$
$$p_{i,s,r,m} \quad = \quad J^T N p_{i,p,r,m} \qquad (3)$$
$$p_{i,p,r,m} \quad = \quad JN p_{i,s,r,m-1} \qquad (4)$$

Basic matrix arithmetic shows that $p_{i,s,l,1}$ is simply $p_{\text{left}}(.|w_i)$ and $p_{i,s,r,1}$ is $p_{\text{right}}(.|w_i)$.

For higher orders, the principle underlying Eq.s 1–4 is that when moving from left to right, we use p_{right} (that is, $J^T N$), the conditional distribution that characterizes right neighbors; when moving from right to left, we use p_{left} (that is, JN), the conditional distribution that characterizes left neighbors. This is graphically shown in Fig. 1.

As illustrated by Fig. 1, the underlying graph for $p_{i,s,r,m}$ and $p_{i,p,r,m}$ is a weighted bipartite directed graph that connects the vocabulary on the left with the vocabulary on the right. A directed edge from w_i on the left to w_j on the right is weighted with $p_{ww}(w_i w_j)/p_w(w_i)$. A directed edge from w_j on the right to w_i on the left (not shown) is weighted with $p_{ww}(w_i w_j)/p_w(w_j)$.

Eq.s 1–4 define four Markov chains:
$$p_{i,s,l,m} = (JNJ^T N)p_{i,s,l,m-1} \qquad (5)$$
$$p_{i,p,l,m} = (J^T NJN)p_{i,p,l,m-1} \qquad (6)$$
$$p_{i,s,r,m} = (J^T NJN)p_{i,s,r,m-1} \qquad (7)$$
$$p_{i,p,r,m} = (JNJ^T N)p_{i,p,r,m-1} \qquad (8)$$

It is easy to see that p_w is a stationary distribution for Eq. 1–4. Writing $\vec{x}$ for p_w, we have:

$$(JN\vec{x})_i = \sum_{j=1}^{V} \frac{p_{ww}(w_i w_j)}{p_w(w_j)} p_w(w_j) = p_w(w_i) = x_i$$

$$(J^T N\vec{x})_i = \sum_{j=1}^{V} \frac{p_{ww}(w_j w_i)}{p_w(w_j)} p_w(w_j) = p_w(w_i) = x_i$$

Hence, p_w is a solution for Eq.s (5)–(8).

The series converge if JNJ^TN and J^TNJN are ergodic, i.e., if the chain is aperiodic and irreducible (Kemeny and Snell, 1976). Observe that for many simple probabilistic context-free grammars (PCFGs) the series in Eq. 1–4 will *not* converge. For simple PCFGs, the alternation between syntagmatic and paradigmatic distributions is periodic. E.g., if inflected verb forms only occur after nouns and nouns only before inflected verb forms, then the right syntagmatic distributions of nouns will have non-zero activation only for verbs and the right paradigmatic distributions of nouns will have non-zero activation only for nouns, thus preventing convergence.[4]

The key difference between a simple PCFG and natural language is ambiguity and noise. Because of ambiguity and noise, JNJ^TN and J^TNJN are likely to be ergodic – there is always a small non-zero probability that two words can occur next to each other. Ambiguity and noise have the same effect as teleportation for PageRank (Brin and Page, 1998) in the sense that we can jump from each word to each other word with non-zero probability.

Assuming that the Markov chains are ergodic, all four converge to p_w: $p_{i,p,r,\infty} = p_{i,p,l,\infty} = p_{i,s,r,\infty} = p_{i,s,l,\infty} = p_w$, for $1 \le i \le V$.

Thus, in this formalization, given enough iterations, syntagmatic and paradigmatic distributions of words eventually all become identical with the prior distribution p_w. This is surprising because linguistically and computationally syntagmatic and paradigmatic relations are fundamentally different.

However, on closer inspection, we observe that limiting the number of iterations is often beneficial when computing solutions to a problem iteratively. E.g., the expectation-maximization algorithm is often stopped early because results close to convergence are worse than results obtained after a small number of iterations. From the point of view of modeling human language acquisition, early stopping is perhaps also more realistic since humans are unlikely to perform a large number of iterations.

---

[4]However, non-ergodicity of JN does not imply non-ergodicity of JNJ^TN and J^TNJN, so Eq. (5)–(8) can converge even for non-ergodic JN.
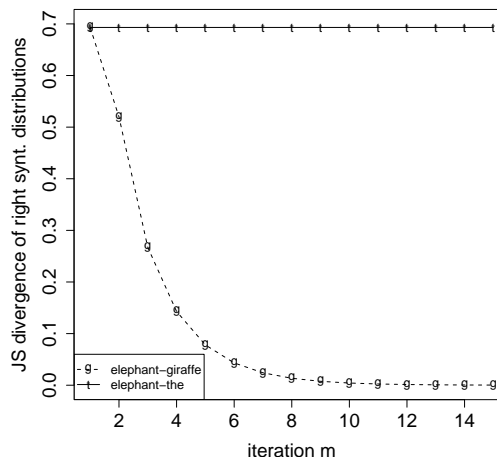

Figure 2: The distance between *elephant* and *giraffe* (measured by the Jensen-Shannon divergence) is accurately represented after a number of iterations. The words *elephant* and *the* retain their large distance.

Example 1. For the following matrix J

$$\begin{pmatrix} & w_1 & w_2 & w_3 \\ w_1 & 82/1002 & 77/1002 & 112/1002 \\ w_2 & 90/1002 & 18/1002 & 107/1002 \\ w_3 & 99/1002 & 120/1002 & 297/1002 \end{pmatrix}$$

we get $p_{1,s,r,1} = (0.31, 0.28, 0.41)$ by computing the product $J^TNp_{1,p,r,1}$. E.g., $p_{1,s,r,1}(w_2) = p_{ww}(w_1w_2)/p_w(w_1) \cdot 1.0 = 77/(82 + 77 + 112) \approx 0.28$.

By iteration $m = 4$, the series $p_{i,s,r,m}$ (Eq. (7)) and $p_{i,p,r,m}$ (Eq. (8)) have converged to:
$$p_{i,s,r,m} = p_{i,p,r,m} = (0.2704, 0.2145, 0.5149)$$
for all three words w_i. One can easily verify that this is p_w. E.g., $p_w(w_1) = (82 + 90 + 99)/1002 = (82 + 77 + 112)/1002 \approx 0.27045$.

Example 2. We computed 15 iterations of syn/para distributions for the corpus: *The giraffe ran. An elephant fell. The man ran. An aunt fell. The man slept. The aunt slept.* Fig. 2 shows that the distance between the right syntagmatic distributions of *elephant* and *giraffe* is large for $m = 1$. The reason is that the two words have no right neighbors in common. The right neighbors of the two words are *ran* and *fell*. Although *ran* and *fell* have no left neighbors in common, their left neighbors have a right neighbor in common: the word *slept*. This indirect similarity information is exploited to deduce by iteration

15 that the two words are very similar with respect to their right syntactic context. In contrast, no such inference, even a very indirect one, is possible for the right contexts of *elephant* and *the*. Consequently, the distance between the two distributions remains high and unchanged with higher iterations.

In this case, the Markov chain is not ergodic and the syntagmatic and paradigmatic series (Eq.s (5)–(8)) do not converge to p_w.

5 Experimental evaluation

Recall from Section 3 that our evaluation task is to discriminate sentences that exhibit local coherence from those that do not; that sentences are represented as sequences of half-words; that syntactic coherence of a sentence is defined as all subsequences of a given length n exhibiting local coherence; and that a subsequence is locally coherent if its distance from a sequence in memory is less than θ.

These definitions can be applied to the graph model as follows. A left half-word is a left syntagmatic (or paradigmatic) distribution and a right half-word is a right syntagmatic (or paradigmatic) distribution. We compute the distance of two half-words either as the Jensen-Shannon (JS) divergence (Lin, 1991) or as $(1 - \cos(\alpha))$. JS divergence is more appropriate for the comparison of probability distributions. But the cosine is more efficient when a sparse vector is compared to a dense vector.[5] We therefore employ the cosine for the compute-intensive experiments in Section 6.

The baseline representation is the categorical representation proposed by Redington et al. (1998). A difficulty in replicating their experiments is that they use hierarchical agglomerative clustering (HAC), which eventually agglomerates all words in a single category. To circumvent the need for a stopping criterion, we represent each word as the temporal sequence of clusters it occurred in during agglomeration and define the distance of two words as the agglomeration step in which the two words are joined in a cluster. E.g., given the agglomeration sequences $\{1\}, \{1,2\}, \{1,2,4\}, \{1,2,3,4\}$ for w_1 and $\{4\}, \{4\}, \{1,2,4\}, \{1,2,3,4\}$ for w_4, the distance

between w_1 and w_4 is 3 since they are joined in step 3 when cluster $\{1,2,4\}$ is created.

For both graded (graph-theoretic) and categorical (cluster-based) representations, we need to set the parameter θ that is the boundary between locally coherent and locally incoherent sentences. This parameter gives rise to a precision-recall tradeoff. A small θ will impose strict requirements on which sequences in memory match, resulting in false negative decisions for local grammaticality. A large θ will incorrectly judge many locally incoherent sequences to be grammatical.

We will pick the optimal θ in both cases. For categorical representations, this amounts to selecting the HAC dendrogram with optimal performance. The experiment below evaluates whether grammatical and ungrammatical sentences are well separated by the proposed measure.[6]

Experiment on CHILDES. We used the well-known CHILDES database (MacWhinney, 2000), a corpus of conversations between young children and their playmates, siblings, and caretakers. In order to avoid mixing varieties of English (e.g., British English vs. American English), we selected the largest homogeneous subcorpus of CHILDES, the Manchester corpus. It contains roughly 350,000 sentences and 1.5 million words. This is a conservative estimate of the amount of child-directed speech a child would receive annually (Redington et al., 1998). All names in the corpus (i.e., all capitalized words) were replaced with a special word "_n_". A boundary symbol "_b_" was introduced to separate sentences. The representation of the corpus is then a concatenation of all its sentences. The vocabulary consists of $V = 8601$ words.

Construction of the evaluation set. We tested the ability of the two models to distinguish locally coherent vs. incoherent sentences by selecting 100 *unattested* sentences from the corpus, which were not used to train the model. We only selected unattested sentences that were not a substring of a sentence in the training corpus since, presumably, any substring of a sentence in the training corpus is locally coherent. A further constraint was that the

---

[5]This is so because, when computing the cosine, we can ignore all dimensions where one of the two vectors has a zero value.

[6]This evaluation of "separation" is not directly an evaluation of classification performance, but more similar to an evaluation of ranking using AUC or an evaluation of clustering using a measure like purity.

unattested sentence was not allowed to contain a word that did not occur in the training corpus, the rationale being that we want to address the problem of local coherence for known words only since unknown words present special challenges. Finally, we ensured that each unattested sentence contained a word that occurred in only one sentence type in the training corpus. In early experiments, we found that local grammatical inference for frequent words is easy as there is redundant evidence available that characterizes legal syntactic environments for frequent words. Since rare words are a key challenge in syntactic acquisition, we only selected sentences as unattested sentences that contained at least one rare word (where a rare word is defined as a word that occurs once in the training set).

100 ungrammatical sentences were generated by randomly selecting and concatenating words from the vocabulary. Ungrammatical sentences were matched in length to unattested sentences, so that both sets contained the same number of sentences of a given length. As with unattested sentences, ungrammatical sentences that were substrings of sentences in the training corpus were eliminated. As there are many more infrequent words than frequent words in the vocabulary, the construction ensured that, as with unattested sentences, infrequent words were overrepresented in ungrammatical sentences.

To summarize, our setup consists of 348,463 training sentences, 100 unattested grammatical sentences and 100 ungrammatical sentences.

The task of discriminating the 100 unattested from the 100 ungrammatical sentences cannot be solved perfectly as CHILDES contains ungrammatical sentences, a few of which were randomly selected as unattested sentences (e.g., *yes pleas*, which is missing the final letter). Similarly, one or two of the automatically generated ungrammatical sentences were actually grammatical.

Since the test set does not consist of a random sample of sentences, performance on the test set is not a direct indicator of the percentage of sentences that the model can correctly discriminate in a child's typical input. A large proportion of sentences in child input are simple 1-word, 2-word, and 3-word sentences that even simplistic models can evaluate with high accuracy. However, the test set is appropriate for a comparative evaluation of graded and
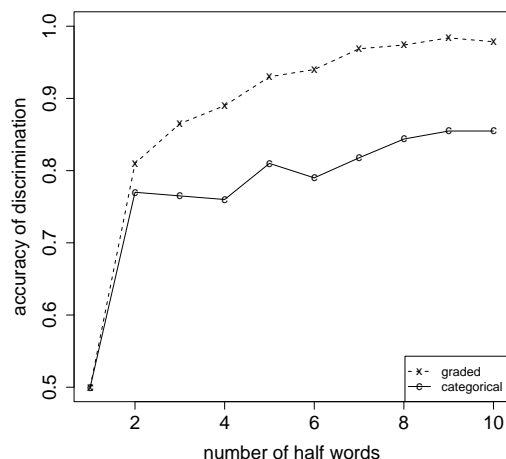

Figure 3: Accuracy of discrimination between grammatical and ungrammatical sentences for graded and categorical representations.

categorical syntactic representations in language acquisition, which is one of the goals of the paper. Difficult sentences (those with rare words and greater length) are overrepresented in the test set as the discrimination of short sentences containing only frequent words can easily be done by simplistic models. Thus, a test set of "easy" sentences would not distinguish good models from bad models.

Discrimination experiment. In order to train the graph model, the entries of matrix J were estimated using maximum likelihood based on the training corpus. $p_{i,s,l,1}$ and $p_{i,s,r,1}$ were then computed for all 8601 words. Replicating (Redington et al., 1998), the most frequent 1000 words were clustered (using single-link HAC, Manning and Schütze (1999)). For each remaining word w, the closest neighbor w' in the 1000 most frequent words was determined and w was then assigned to the cluster of w'.

Fig. 3 shows the performance of graded and categorical representations for different subsequence sizes n. To compute the accuracy for each n, the θ with optimal discrimination performance was chosen (for both graded and categorical).

For a subsequence of size $n = 1$, the performance is 0.5 in both cases since the 200-sentence test set does not contain unknown words. So for every half-word, there is a sequence of one half-word in the training corpus with distance 0. Thus, all sentences

923

get the same local coherence scores, both for graded and categorical representations.

This argument does not apply to $n = 2$ since we earlier defined a sentence to be locally coherent if all of its subsequences are coherent. While subsequences of 2 half-words that are part of the *same* word have local coherence score 0, this is not true of subsequences of 2 half-words that are part of *different* words, e.g., the subsequence $<black_r, dog_l>$ in *black dog*. If *black dog* does not occur in the training set, then its local coherence score is > 0.

The main result of the experiment is that except for n=1 ($p = 1$) and n=2 ($p = 0.39$) the differences between categorical and graded representations are significant (χ^2 test, $p < 0.05$ for $3 \leq n \leq 10$). This is evidence that graded representations are more accurate when determining local syntactic coherence and grammaticality than categorical representations.

The experimental results demonstrate that, for syntagmatic distributions of order 1, graded representations discriminate locally coherent vs. incoherent sentences better than categorical representations. We attribute this to the ability of exemplar theory to incorporate rich context information into discrimination decisions. This is of particular importance for ambiguous words. Categorical representations of ambiguous words are problematic because they are either too similar or not similar enough to the two alternatives. E.g., if a word with a verb/noun ambiguity is represented as one of the alternatives, say, as a verb, then subsequences containing its noun use will no longer be similar to other subsequences with nouns. If a special conflation category noun/verb is introduced, then we are faced with the same problem: subsequences containing the noun/verb category are not similar to subsequences containing either non-ambiguous verbs or non-ambiguous nouns.

6 Higher-order distributions

The main motivation for higher-order distributions is that syntagmatic vectors of order 1 do not perform well for some infrequent words. In the elephant/giraffe example above, the distance between the two words is close to maximum for order 1 representations because each occurs only once, in entirely different contexts. As we showed in Fig. 2, higher-order representations address this problem because
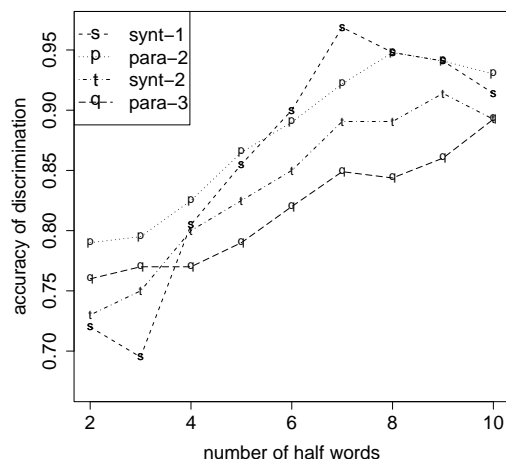

Figure 4: Accuracy of discrimination between grammatical and ungrammatical sentences of the exemplar-based method for different orders. Key: synt = syntagmatic, para = paradigmatic; s is of order 1; p and t are of order 2; q is of order 3.

they exploit indirect evidence about the syntactic properties of words.

To evaluate higher-order representations on CHILDES, we used the same setup as before, but computed several additional iterations. We also limited the experiments to a subset consisting of 60,000 words of the Manchester corpus. It contains only V=1666 different words, which reduces the storage requirements for the syn/para distributions (which is $2 \cdot V^2$ for each order) and the cost of the matrix multiplications. We also used $(1 - \cos(\alpha))$ instead of JS divergence as distance measure.

The results of the experiment are shown in Fig. 4. Higher-order representations are clearly superior for short subsequences, especially for $n = 2$ and $n = 3$ (and up to 5 half-words when comparing synt-1 and para-2). However, for long subsequences, there is no consistent difference between the syntagmatic distribution of order 1 (synt-1) and higher order distributions. Apparently, the generalized information available in higher orders is not helpful in local grammatical inference if long contexts are considered.

We were surprised that the best-performing distribution for short sequences is para-2 (paradigmatic distribution of order 2), not a higher order distribution. E.g., para-3 performs worse than para-2.

924

We would expect the performance to decrease with higher order eventually since the distributions converge towards p_w. The fact that this happens so early in this experiment merits further investigation.

7 Related work

Data-oriented parsing (Bod et al., 2003) shares basic assumptions about linguistic inference with exemplar-based theory, but it does not model or use the similarity between input and stored exemplars. Previous work on exemplar theory in syntax (Abbot-Smith and Tomasello, 2006; Bybee, 2006; Hay and Bresnan, 2006) has not been computational or formal. Previous work on non-categorical representations of words has viewed these representations as an intermediate step for arriving at categorical parts of speech (Redington et al., 1998; Schütze, 1995; Clark, 2003). Consequently, all of these papers evaluate their results by comparing induced categories to gold-standard parts of speech.

Redington et al. (1998) did not find a difference in categorization accuracy between simple syntagmatic representation and those using non-adjacent words.

The BEAGLE model (Jones and Mewhort, 2007), and related work (Sahlgren et al., 2008), merges co-occurrence information and word order information into a single composite vector through a process of vector convolution. Our model differs in that it explicitly captures the recursive relationship between the orders in a unified framework.

Previous graph-theoretic work (Biemann, 2006) uses order 1 representations. Several papers have looked at higher-order representations, but have not examined the equivalence of syn/para distributions when formalized as Markov chains (Schütze and Pedersen, 1993; Lund and Burgess, 1996; Edmonds, 1997; Rapp, 2002; Biemann et al., 2004; Lemaire and Denhière, 2006). Toutanova et al. (2004) found that their graph model of predicate argument structure deteriorated after a small number of iterations of the random walk, similar to our findings.

8 Conclusions and Future Work

In this paper, we have presented a graph-theoretic model of the acquisition of lexical syntactic representations and a new exemplar-based evaluation of lexical syntactic acquisition. When applied to the CHILDES corpus, the evaluation shows that the graded syntactic representations learned by the model perform significantly better than previously proposed categorical representations. An initial evaluation of high-order representations showed little improvement over low-order representations.

In future work, we intend to investigate the influence of noise and ambiguity on the quality of the representations in order to characterize when higher order representations improve generalization and exemplar-theoretic inference. We also want to address that the model as it currently stands is trained under the false assumption that the training input is grammatical. Ungrammatical test input which matches a learned ungrammatical sequence will be deemed grammatical. Future work will examine how to best treat this challenge, e.g., by using an estimation of density instead of the simplistic "1 nearest neighbor" distance used here.

The most important future work concerns class-based language models. The cognitive-linguistic tradition we have mainly addressed in this paper has focused on the task of learning traditional parts of speech and has usually not discussed the relevance of language models to acquisition. If, as we have argued, instead of learning traditional parts of speech the focus should be on performance in particular language processing tasks (like grammaticality judgments), then language models are the natural competing account that we must compare our work to. Of particular relevance are class-based language models (e.g., (Saul and Pereira, 1997; Brown et al., 1992)). In ongoing work, we are attempting to show that the exemplar-theoretic model performs better on grammaticality judgments than class-based language models.

References

Abbot-Smith, Kirsten and Michael Tomasello. 2006. Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, 23:275–290.

Biemann, Chris, Stefan Bordag, and Uwe Quasthoff. 2004. Automatic acquisition of paradigmatic relations using iterated co-occurrences. In *LREC*.

Biemann, Chris. 2006. Unsupervised part-of-speech tagging employing efficient graph clustering. In *ACL*.

Bod, Rens, Remko Scha, and Khalil Sima'an. 2003. *Data-Oriented Parsing*. CSLI Publications.

Brin, Sergey and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *WWW*, pages 107–117.

Brown, Peter F., Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4):467–479.

Bybee, Joan L. 2006. From usage to grammar: The mind's response to repetition. *Language*, 82:711–733.

Christiansen, Morten, Luca Onnis, Padraic Monaghan, and Nick Chater. 2004. Happy endings in language acquisition. In *AMLaP*.

Church, Kenneth, Patrick Hanks, Donald Hindle, William Gale, and Rosamund Moon. 1994. Lexical substitutability. In Atkins, B.T.S. and A. Zampolli, editors, *Computational Approaches to the Lexicon*. OUP.

Clark, Alexander. 2003. Combining distributional and morphological information for part of speech induction. In *EACL*, pages 59–66.

de Saussure, Ferdinand. 1962. *Cours de linguistique générale*. Payot, Paris. Originally published in 1916.

Edmonds, Philip. 1997. Choosing the word most typical in context using a lexical co-occurrence network. In *ACL*, pages 507–509.

Freudenthal, Daniel, Julian Pine, and Fernand Gobet. 2004. Resolving ambiguities in the extraction of syntactic categories through chunking. In *ICCM*.

Gleitman, Lila and Elissa Newport. 1995. The invention of language by children: Environmental and biological influences on the acquisition of language. In Gleitman, Lila and Mark Liberman, editors, *Language: An invitation to cognitive science*. MIT Press, 2nd edition.

Gleitman, Lila. 1990. The structural sources of verb meanings. *Language Acquisition*, 1:3–55.

Goldinger, Stephen D. 1997. Words and voices—perception and production in an episodic lexicon. In (Johnson and Mullennix, 1997).

Hay, Jennifer and Joan Bresnan. 2006. Spoken syntax: The phonetics of giving a hand in New Zealand English. *The Linguistic Review*, 23.

Johnson, Keith and John W. Mullennix, editors. 1997. *Talker Variability in Speech Processing*. Academic Press.

Johnson, Keith. 1997. Speech perception without speaker normalization. In (Johnson and Mullennix, 1997).

Jones, Michael N. and Douglas J.K. Mewhort. 2007. Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114:1–37.

Kemeny, John G. and J. Laurie Snell. 1976. *Finite Markov Chains*. Springer, New York.

Lemaire, Benoit and Guy Denhière. 2006. Effects of high-order co-occurrences on word semantic similarity. *Behaviour, Brain & Cognition*, 18(1).

Lin, Jianhua. 1991. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory*, 37(1):145–151.

Lund, Kevin and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, 28:203–208.

MacWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Boston, MA.

Mintz, Toben H. 2002. Category induction from distributional cues in an artificial language. *Memory & Cognition*, 30:678–686.

Pierrehumbert, Janet. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In Bybee, Joan and Paul Hopper, editors, *Frequency and the Emergence of Linguistic Structure*, pages 137–157. Benjamins.

Rapp, Reinhard. 2002. The computation of word associations: comparing syntagmatic and paradigmatic approaches. In *Coling*.

Redington, Martin, Nick Chater, and Steven Finch. 1998. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4):425–469.

Sahlgren, Magnus, Anders Holst, and Jussi Karlgren. 2008. Permutations as a means to encode order in word space. In *CogSci*.

Saul, Lawrence and Fernando Pereira. 1997. Aggregate and mixed-order markov models for statistical language processing. In *EMNLP*, pages 81–89.

Schütze, Hinrich and Jan Pedersen. 1993. A vector model for syntagmatic and paradigmatic relatedness. In *UW Centre for the New OED and Text Research*.

Schütze, Hinrich, Michael Walsh, Travis Wade, and Bernd Möbius. 2007. Towards a unified exemplar-theoretic model of phonetic and syntactic phenomena. In *CogSci, Poster Session*.

Schütze, Hinrich. 1995. Distributional part-of-speech tagging. In *EACL*, pages 141–148.

Toutanova, Kristina, Christopher D. Manning, and Andrew Y. Ng. 2004. Learning random walk models for inducing word dependency distributions. In *ICML*.

Wade, Travis, Grzegorz Dogil, Hinrich Schütze, Michael Walsh, and Bernd Möbius. 2008. Syllable frequency effects in a context-sensitive segment production model. Submitted.

926