Distributional Learning of Syntactic Categories

Kevin Liang, Diana Marsala, and Charles Yang

1. Introduction

Syntactic categories are the foundation of grammar but their status in linguistic theory and language acquisition remains an open question. Indeed, there is still ongoing debate as to whether syntactic categories are universally attested in the world's languages (Pinker, 1984; Comrie, 1989). In this paper, we will continue to use terms such as nouns and verbs as a matter of convenience, but we are open to the possibility that syntactic categories are the result of distributional regularities in language (Harris, 1955; Chomsky, 1955; Maratsos & Chalkley, 1980), and we believe our approach lends credence to that possibility. At the same time, the problem of syntactic category learning remains. Even if syntactic categories are innate and universal, an English-learning child still has to learn that *cat* is a noun, *see* is a verb, but *jump* may be both.

According to the semantic bootstrapping hypothesis (Macnamara, 1982; Pinker, 1984), children rely on the semantic content of words, which may be available from observation, to establish syntactic categories. For example, they may know that persons, animals, and objects are nouns, actions are verbs, spatial relations are prepositions, properties are adjectives, etc. The semantic bootstrapping hypothesis gains support from findings that even infants have a rich understanding of the world, including objects, event structures, causation, intentionality, and other conceptual categories that are systematically mapped to syntactic categories in language; see Carey (2009) for review. Furthermore, caretaker speech appears to facilitate semantic bootstrapping. For example, Rondal and Cession (1990) find that nearly all persons and objects are described by nouns and nearly all actions and states are described by verbs. However, as its name entails, the semantic bootstrapping hypothesis is only a start. The meanings of many, perhaps most, words cannot be readily learned from observation, but may rely on the development and use of a formal grammatical system that is composed of syntactic categories (Gleitman, 1990).

Indeed, very young children already know a great deal about the formal grammatical system of their language. For example, infants use phonological cues to categorize determiners (Shi, Werker, & Cutler, 2006; Shi & Melançon, 2010)

* Kevin Liang, University of Pennsylvania, kevliang@sas.upenn.edu. Diana Marsala, University of Pennsylvania. Charles Yang, University of Pennsylvania, charles.yang@ling.upenn.edu

 Proceedings of the 46th annual Boston University Conference on Language Development, ed. Ying Gong and Felix Kpogo, 442-455. Somerville, MA: Cascadilla Press.

which can then be used as distributional cues for other syntactic categories. Additionally, young children's ability for formal pattern learning is well documented (Saffran, Aslin, & Newport, 1996; Gary F Marcus, Vijayan, Rao, & Vishton, 1999; Gómez & Lakusta, 2004) and can be recruited for categorization in the absence of meaning (Toben H Mintz, Newport, & Bever, 2002; LouAnn Gerken, Wilson, & Lewis, 2005). An especially useful cue in syntactic categorization is the notion of a *frequent frame* (Toben H. Mintz, 2003), which is defined as words that frequently co-occur with one word intervening. For example, in the utterance "the cat on the box", [*the _ on*] is a frame that contains the word "cat" and, with very high accuracy, other nouns in the language. Frequent frames can thus be used to establish equivalent classes, which is the function of syntactic categories.

In this paper, we propose a model of syntactic category learning, the *Category Frame Learner* (CFL), which combines the core elements of semantic bootstrapping and distributional learning. The central component of CFL is the Tolerance Principle (TP; Yang, 2016), an independently motivated principle that provides precise conditions for productive linguistic rules. The TP enables the learner to combine and generalize frequent frames, which are defined in terms of specific lexical items, into category frames, which are defined in terms of broader (syntactic) categories. As a result, if the child knows just a handful of words whose meanings align closely with their syntactic categories (e.g., a few concrete objects as nouns), CFL can form category frames from the lexical frames, greatly expanding their applicability. Crucially, unlike other models of syntactic category acquisition, CFl is parameter-free. We show that CFL provides accurate syntactic categorization on corpora of child-directed English, Mandarin, French, and especially German, a language for which frequent (lexical) frames have been shown to be ineffective (Stumper, Bannard, Lieven, & Tomasello, 2011). We also show that the CFL has the flexibility to include other distributional cues such as morphology (Brown, 1957), further increasing categorization accuracy.

2. Background

2.1. Distributional Cues for Syntactic Category

While frequent frames can produce highly accurate word classes (Toben H. Mintz, 2003), they lead to a proliferation of syntactic categories. For example, [*the _ on*] is almost guaranteed to find nouns, but not all nouns appear in this very specific lexical frame: many will appear in frames that are similar (e.g., [*a _ in*]) or quite different (e.g., [*the _ is*]). A proliferation of syntactic categories ensues. One may argue that such an outcome is not unwelcome as there are many subclasses of nouns with subtle syntactic and semantic differences (e.g., animacy, count/mass). However, an increase in the number of syntactic categories rapidly increases the number of potential syntactic combinations and does not help with the overall task of language acquisition. In addition, there is compelling evidence that children do in fact form equivalent classes of words similar to syntactic categories in adult grammar; see Dye, Kedar, and Lust (2019) for review.

Therefore, lexical frames must be combined to form more abstract frames. If the child observes *cat* in both [*the* _ *on*] and [*a* _ *in*], they should seek to establish additional regularity. In particular, if they also know that *the* and *a* are D(eterminers) and *on* and *in* are P(repositions), they may conclude that [D _ P], a *category frame*, is effective for categorizing nouns (or whatever category *cat* belongs to). Both children and adults have been shown to make such generalization in artificial language studies (Reeder, Newport, & Aslin, 2013; Hall, Van Horne, & Farmer, 2018). However, no effective strategy has been proposed for frame generalization: for example, a recursive procedure for combining frames (Chemla, Mintz, Bernal, & Cristophe, 2009) actually produced worse results than lexical frames.

The key question for frame generalization is in fact central to the study of language acquisition: How do children form categorical generalizations from lexically specific patterns in the input data? We now turn to the Tolerance Principle (TP), a recent theory of how linguistic generalizations emerge.

2.2. The Tolerance Principle

The TP asserts:

(1) Let a rule R be defined over a set of N items. R is productive if and only if e, the number of items not supporting R, does not exceed θ_N:

$$e \leq \theta_N = \frac{N}{\ln N}$$

If e exceeds the threshold θ_N, a learner does not generalize the rule R and memorizes the properties of all N items.

The TP has been applied to many problems in language acquisition, variation, and change. It also receives support from artificial language learning experiments (Schuler, Yang, & Newport, 2016; Koulaguina & Shi, 2019; Emond & Shi, 2021) with precisely controlled conditions. In particular, we use the TP to identify category frames, as productive generalizations over lexical frames.

3. The Category Frame Model

3.1. Semantic Seeds

The CFL model is initialized with a small set of frequent and semantically salient words, known as *seeds*, along with their syntactic categories. We will again refer to the categories as nouns, verbs, etc. without assuming they are the substantive universals of language. Rather, we assume that children understand the seeds as members of semantic or conceptual categories (e.g., animals, objects, actions), available innately or otherwise. Their task is to discover and generalize the distributional properties of these words, using the lexical and category frames

discussed earlier. In this process, the initially semantic categories develop into formal categories (e.g., "noun"), which will include words whose content (e.g., "neutron") bears little resemblance to the seeds (e.g., "apple") provided in the bootstrapping stage except for their formal distributional similarities. There are 7 or 8 syntactic categories, each of which has a handful seeds; see Experiments below for details.

3.2. Lexical Frames and the Lexicon

The CFL model is provided with a training corpus and a set of seed words. It maintains a lexicon ($\mathcal{L}$) which records the category labels of words that are deemed sufficiently reliable (see below). It derives a set of frames ($\mathcal{F}$) for syntactic categorization. Here we focus on lexical frames of the form [a _ b], where the items surrounding "_" are words (Toben H Mintz, 2003), denoted by lower case letters.

Initially $\mathcal{L}$ only contains the seeds (and their labels) and the set of frames $\mathcal{F}$ is null. The CFL first identifies lexical frames that are deemed trustworthy, in a sense to be made clear, before attempting to generalize them to category frames (section 3.3).

For each utterance in the training data, the model scans from left to right and identifies all lexical frames of the form [a x b]. It moves into action if and only if the word x is in $\mathcal{L}$ (i.e., has been learned with a label) or [a _ b] is in $\mathcal{F}$, i.e., a trusted (lexical) frame. A lexical frame becomes trusted when it has correctly labeled words that are already in $\mathcal{L}$ a sufficiently large number of times, in a scoring scheme described below.

If a word in the frame (i.e., a or b) is in $\mathcal{L}$ and [a _ b] does not exist in $\mathcal{F}$, then a new frame is created. For instance, if the context is [*the* red *dog*] where "red" is in $\mathcal{L}$ as an ADJ(ective), then the frame [*the* ADJ *dog*] is created with a score of 1.

For simplicity, the score of a frame is handled by a counter, which increases by 1 if the frame succeeds to categorize a known word in $\mathcal{L}$ and decreases by 1 if it fails. For example, suppose that the input utterance contains the string "the very dog that barked the loudest" from which the frame [*the* very *dog*] is available. Suppose further that $\mathcal{L}$ contains the word "very" as an adverb. The frame [*the* ADJ *dog*] will be penalized by decreasing its score by 1. This reward and penalty scheme is motivated by successful applications of Reinforcement Learning to grammar and word learning (Yang, 2002; Stevens, Gleitman, Trueswell, & Yang, 2017)) although other updating functions are also possible. If the score of a frame falls below τ_f, it is removed from $\mathcal{F}$, i.e., no longer trusted. In other words, the list of trusted frames is dynamic.

Clearly, the quality of lexical frames depends on the words in $\mathcal{L}$ having been correctly identified; otherwise there will be a proliferation of inaccurate frames. Recall that $\mathcal{L}$ is initialized as the set of pre-specified seeds, which are presumed to be correct. It is therefore important to ensure that when new words and their

labels are added to $\mathcal{L}$, they must also be of high quality. This again is done by a scoring scheme – for words, rather than frames.

A word will be assigned a label only if it has been tagged by a trusted frames in $\mathcal{F}$. Then, the score for that label increases by 1. It is possible for a word in $\mathcal{L}$ to be tagged a different label (by other trusted frames in $\mathcal{F}$). In that case, the new label is initialized with the score of 1; the existing labels are penalized by decreasing its score by a penalty parameter set to 0.75 throughout. The setting of the penalty to be smaller than the reward is motivated by findings in cross-situational word learning: previously hypothesized but subsequently disconfirmed hypotheses do not entirely vanish from memory (Köhne, Trueswell, & Gleitman, 2013; Stevens et al., 2017). If the score of a label exceeds a threshold τ_w, it is entered into $\mathcal{L}$. Words with multiple labels are thus ambiguous across syntactic categories (e.g., *jump*), and the scores for the labels represent usage preferences. The fact that the penalty value is smaller by the reward value enables multiple labels to be learned for a single word (i.e., syntactic category ambiguity). Like frames, a word label may be ejected from $\mathcal{L}$ if its score falls below τ_w.

The frame set $\mathcal{F}$, which is initially empty, gradually begins to include trusted lexical frames that are acquired from the seed words and their labels. These lexical frames, once trusted, will be able to tag additional words and gradually enlarge $\mathcal{L}$, which in turn produces additional trusted lexical frames in an iterative process. We now turn to discuss how lexical frames give rise to category frames.

3.3. From Lexical Frames to Category Frames

The critical component of the CFL model is the generalization of category frames from lexical frames via the TP. Both *partial* - those with a syntactic category on one side of the frame only - and *full* - those with syntactic categories on both sides of the frame - category frames are created.

For each syntactic label X, consider all the trusted lexical frames of the form [a X b] in $\mathcal{F}$: suppose that there are N such frames. Suppose that a has been tagged as A in $\mathcal{L}$ and b as B in $\mathcal{L}$. Consider the partial category frames of the form [A _ b] and [a _ B] and the full category frame [A _ B], where the uppercase letters A and B denote syntactic categories. For each of these, the number of frames that take on that form is greater than $N - N/\ln N$, generalize the frame, which is subsequently entered in the set of frames in $\mathcal{F}$.

For example, consider a situation where there are $N = 10$ trusted lexical frames that tag nouns with the following distribution: 7 of the form [*the* _ V] (where V represents seven different verbs in $\mathcal{L}$), 2 of the form [*blue* _ V] (where V represents two different verbs in $\mathcal{L}$), and 1 of the form [*mom* _ *likes*]. The child would generalize the frame [*the* _ V] to tag nouns since it occurs 7 times, exceeding the requisite Tolerance Principle threshold for productivity (10/ln10=4).

With the introduction of category frames, multiple frames may be applicable to a new word in the input. Recall that learning only takes place if the word is in $\mathcal{L}$ with a label. Among all applicable frames, the order of precedence is lexical

frame first, followed by partial and then full category frames. This process embodies the so-called Elsewhere Condition (Anderson, 1969) in linguistics, or perhaps a more general cognitive principle that favors specificity. The model searches for the most specific frame that predicts the word label correctly and rewards that frame; all frames that lead to it (and have failed) are penalized.

3.4. Morphological Cues for Syntactic Categories

The distributional information available to children is not limited to words; grammatical morphemes such as *-ing* appear highly effective for syntactic categories as has been demonstrated in a wide range of studies (LouAnn Gerken, Landau, & Remez, 1990; Santelmann & Jusczyk, 1998; Shi, Cutler, Werker, & Cruickshank, 2006; Soderstrom, Blossom, Foygel, & Morgan, 2008; Van Heugten & Johnson, 2010).

We create a method for automatically extracting the morphological cues for syntactic categories. After training, the CFL model learns a lexicon $\mathcal{L}$ with words and their label scores. Word labels with high scores are those that have been repeatedly confirmed, and as we will see below, highly accurate. Among these, we look for transparently related word pairs within each syntactic category. In English, we always find words such as *cat*, *cats*, *boy*, *boys*, etc. in the category noun, and words such as *eat*, *eating*, *kiss*, *kissing* in the category verb. We extract the morphemes, by simple string comparison, that relate multiple such pairs and also mostly *unambiguously* identifies a single syntactic category. Thus, the German ending *-en*, which relates verb pairs (as past participle) but also noun pairs (as plural), is not extracted. In this way, we consistently identify the following morphemes for syntactic categories: English - *V/-ing, -ed, -en*, *N/-s*, *ADJ/-est*; French - *V/-ir, -er, -re, -ez, -ons, -ent*, *ADV/-ment*; German - *V/-est,-et*, *ADJ/-er, -es, -em*. These morphemes are then added to $\mathcal{F}$ as frames: words with such endings are immediately tagged as the corresponding syntactic category. In the present implementation, the morphological cues are extracted after training; it is possible to integrate the process during the training phase as well.

4. Experiments

For each experiment, we held out 10% of the data as a testing set. All results were averaged over 10-fold cross-validation. The results of these experiments were compared to a baseline that simply tagged the provided seed words for each corpus. During the testing phase, only the frames in $\mathcal{F}$ were used for tagging. The threshold τ_w in training $\mathcal{L}$ no longer applies and words were immediately tagged with a label. If no frame in $\mathcal{F}$ is applicable but the word is in $\mathcal{L}$, then its most likely label is produced. Otherwise, the most frequent label in $\mathcal{L}$ is produced (which is always a noun across all languages in our experiment).

4.1. Corpora

The experiments were carried out on child-directed input from CHILDES (MacWhinney, 2000). For English, we used one large corpus - the Manchester corpus (Theakston, Lieven, Pine, & Rowland, 2001). This corpus is a logitudinal study of British-English speaking children. The children in the study were recorded weekly over the course of a year from the age of 2 to the age of 3. In total, there are 373986 sentences of child-directed speech in the corpus.

To demonstrate the efficacy of the CFL model cross-linguistically, we also chose to evaluate the effectiveness of the model on a combination of Mandarin, French and German CHILDES corpora (references omitted). For Mandarin, French, and German, we had to combine several corpora although their collective size is still smaller. The sizes of the corpora for each language were as follows: Mandarin - 164705 sentences, French - 202262 sentences, and German - 230190 sentences.

Additionally, we also evaluate the CFL model on the Wall Street Journal portion of the Penn Treebank (Mitchell P. Marcus, Marcinkiewicz, & Santorini, 1993) and compare it to the results in Haghighi and Klein (2006). There were a total of 59100 sentences for the English section and 28295 sentences for the Mandarin section. These experiments aim to demonstrate the overall effectiveness of the CFL model compared to significantly more complex models.

For all child-directed corpora, we converted the syntactic categories from CHILDES to 7 basic categories loosely following the Universal Dependency Treebank annotation scheme (Nivre, 2017) in order to facilitate cross-linguistic comparison. These are adjective, adverb, determiner, noun, preposition, pronoun, and verb. For Mandarin, we added a classifier category. Words that are not mapped to these categories, about 5% of all tokens, are kept in the data but do not participate in training or testing.

For each category, we chose a small number of frequent and semantically salient words as seeds. For example, the following 28 seed words used for English were the following - pronoun: *you, we, me*; verb: *come, play, put*; preposition: *on, out, in*; determiner: *this, these*; noun: *baby, car, train, box, house, boy, man, book*; adjective: *big, silly, green*; adverb: *well, very, now*; conjunction: *and, or, but*. We did not find significant performance variation with the lexical choices of the seeds although more seed words helped with the smaller corpora to speed up learning.

Lastly, for the WSJ corpora, we use the same seed list as Haghighi and Klein (2006) of 112 words across 45 POS tags to allow for a better comparison.

CFL has two threshold parameters – τ_f and τ_w,- which determine the entry of words and frames into $\mathcal{L}$ and $\mathcal{F}$. These values are empirical in nature: behavioral studies may determine, for example, the amount of exposure for children to accurately learn the syntactic categories of words. We experimented with the various values and set both to 15 for English and Mandarin, 12 for French corpus, and 6 for German corpus. The choices of parameter values are essentially determined by the varying corpus size of the languages under study. We expect that a single, and relatively high, threshold would work well across languages if we had more child-

directed data (e.g., a few million words), distinguishing the CFL model from other models which require parameter estimation. The CFL model is an online and incremental algorithm and thus highly efficient. For example, training on the largest corpus (English) takes roughly ten minutes on a consumer-level laptop.

4.2. Evaluation

For each experiment, we first calculated the one-to-one accuracy of the model on test data after training. This was obtained by simply dividing the number of words tagged correctly in the testing set by the total number of words. Other scoring methods are possible but one-to-one accuracy is shown to be appropriate when gold standard test data is available as is in the case here.

While overall accuracy is important, we will also focus on words that the model is most confident about, because they would be the earlier words that children learn, and also those that play a critical role in the development of grammar. They should be as closely aligned with the target syntactic categories. To this end, we calculate pairwise precision and recall - typically used to measure the quality of clustering tasks (Christodoulopoulos, Goldwater, & Steedman, 2010).

Pairwise precision and recall are defined using each pair of words tagged. If two words have the same (true) label as well, then the pair is counted as correct if the model assigns the two words the same tag. Conversely, if the two words have different true tags, then the pair is counted as corrected if the model gives the two words different tags. For example, suppose the model produced two categories (*cat*, *dog*, *come*) and (*have*, *go*, *tea*). Each category has three members, and thus three pairwise comparisons. Both categories thus have the precision of 1/3. As for recall, the true categories ought to be (*cat*, *dog*, *tea*) and (*have*, *go*, *come*). There are again three pairs in each category but only 1 is grouped in the same category by the model. Both categories thus have the recall of 1/3 as well.

5. Results

5.1. CHILDES Results

We first report the accuracy of the CFL model compared to the baseline on CHILDES data is shown in Table 1. Results from the addition of morphological features are also reported as CFL + MOR. The baseline accuracy is the accuracy using only the seed words tagged in the test data.

Table 1: CFL accuracies on child-directed input

Model	Eng	Man	Fre	Ger
Baseline	14.7%	24.7%	13.4%	16.4%
CFL	65.9%	62.0%	50.9%	58.4%
CFL + MOR	72.8%	-	59.3%	70.4%
Category Frames Only	64.2%	62.1%	51.5%	62.9%
Lexical Frames Only	73.3%	63.8%	59.5%	66.5%

On the CHILDES corpora, the CFL model achieved a reasonably high accuracy across languages. The high performance on Mandarin, the smallest dataset, is possibly due to the stricter word order in Mandarin. The addition of the morphological features (inapplicable to Mandarin) proved highly effective, leading to significant improvement over the base CFL model alone.

The number of full category frames learned over time in the English CHILDES model is shown in Figure 1, which seem to plateau after 124,000 sentences with 12 frames: there is no effect of over-fitting. This is not entirely surprising given that we have 7 categories which can only lead to a maximum of 49 possible full frames. It is unlikely that a significantly higher number of frames could reasonably be generated by the model since many frames will co-occur with multiple syntactic categories; thus, those frames will not meet the threshold for generalization under the TP. Table 2 lists some samples of the frames generated by the model.

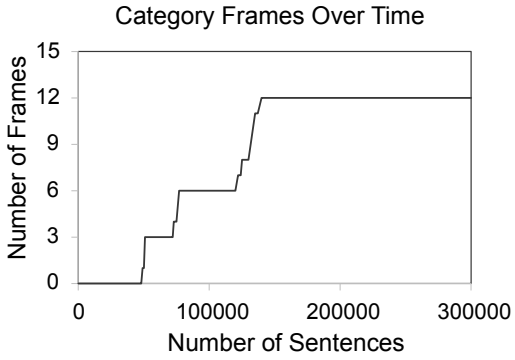

Figure 1: Generalized frames overtime

Table 2: Sample category frames generated

Frame	Tag	Example
[PRO _ PRO]	VERB	did we eat it all up
[DET _ VERB]	NOUN	what does that rooster say
[VERB _ VERB]	CONJ	sit and take a break
[ADV _ NOUN]	ADJ	very nice noise

The benefits of category frames, especially the partial and full category frames, can be observed during the *final* stage of training. By then, approximately 60% of words that had not been labeled before were tagged by a combination of either partial or full category frames. This suggest that the use of the TP in CFL expanded the coverage of syntactic categories considerably beyond the use of lexical frames alone as in previous work.

Next, we calculated pairwise precision, recall, and F1-score across all tokens.

Table 3: CHILDES Pairwise Metrics

Language	Precision	Recall	F1-score
English	0.569	0.788	0.661
Mandarin	0.512	0.777	0.617
French	0.530	0.616	0.570
German	0.498	0.696	0.581

The CFL model was able to achieve a high recall during the testing phase with reasonable precision. This means that the model was able to both label a large majority of words in the testing phase and mostly categorize them correctly.

Importantly, CFL performs even better when only words with high label scores are considered. To achieve high scores, these labels must have been confirmed repeatedly during the training period. The precision on these words is significantly higher as shown in Table 4. These results suggest that the model is able to generate a vocabulary of highly accurate words. These words are likely to be among the earliest words children learn and can be expected to play a critical role in the later stages of language development.

Table 4: Pairwise Metrics on Most Confident Words

Language	Precision
English	0.913
Mandarin	0.867
French	0.846
German	0.704

The most comparable model in the literature is Christodoulopoulos, Roth, and Fisher (2016). However, unlike our model, Christodoulopoulos et al. (2016) uses a variable number of seed words, and it only categorizes nouns and verbs. Additionally, the model was run on the Brown corpus (Brown, 1973) in CHILDES which resulted in slightly different F1 scores when the CFL model was run on it. Despite the simpler task, the Christodoulopoulos et al. (2016) model is only able to achieve an F1-score of 0.471 with 15 seed words (c.f. an F1-score of 0.743 achieved by the CFL model on the Brown corpus).

We are not aware of other comparable models in the cognitive modeling literature. Most previous models are completely unsupervised (Redington, Chater, & Finch, 1998; Toben H. Mintz, 2003; Chemla et al., 2009). It is not clear, however, how these models would benefit from a small number of seed words.

Nevertheless, we provide comparisons to these models to demonstrate the efficacy of the CFL model in Table 5 and Table 6. These approaches use different corpora which gave slightly different performance metrics for our model with Cartwright and Brent (1996) using the Brown corpus (Brown, 1973) and Toben H. Mintz (2003) using a combination of various English-language CHILDES corpora. Although the precision of our model is lower than Cartwright and Brent

(1996) and Toben H. Mintz (2003), the recall is much higher. This means that the CFL model is able to tag a significantly higher proportion of the total words which is reflected in the overall higher F1-score relative to the other models.

Table 5: Brown Corpus Pairwise Metrics

Model	Precision	Recall	F1-score
Cartwright	**0.853**	0.178	0.295
CFL	0.627	**0.912**	**0.743**

Table 6: Mintz Corpora Pairwise Metrics

Model	Precision	Recall	F1-score
Mintz	**0.910**	0.130	0.228
CFL	0.585	**0.712**	**0.642**

5.2. WSJ Results

Finally, we evaluate the CFL model on the Wall Street Journal portion of the Penn Treebank (Mitchell P. Marcus et al., 1993) and compare it to the results in Haghighi and Klein (2006) (H&K) - an important result that had significant impact in (minimally supervised) language learning but was obtained with much more complex optimization techniques. In addition, H&K made use of lexical similarities gathered from This dataset has 45 hand-annotated categories, with 112 seeds spread across them.

The comparisons are shown in Table 7. The baseline accuracy is just from the seeds alone. The PROTO model in Haghighi and Klein (2006) is similar to the CFL model in that it uses three seed words for each syntactic category. However, it also uses spelling features and optimization techniques that would not be feasible for the child learner. The PROTO+SIM model adds on to the PROTO model by also including word context vectors.

Table 7: WSJ Model Accuracies

Model	English	Mandarin
Baseline	42.3%	29.4%
CFL	55.7%	**57.9%**
H&K PROTO	68.8%	39.0%
H&K PROTO+SIM	**71.5%**	57.4%

On the WSJ dataset, our model for English is not able to perform better than the PROTO model in Haghighi and Klein (2006). However, when our model is run on Mandarin, the performance is significantly higher despite the additional computational complexity of the Haghighi and Klein (2006) models and the use of features inaccessible to the child learner. Although high accuracy across all to-

kens is not necessarily needed for the child learner, it does demonstrate the overall efficacy of our model.

6. Conclusion

In this paper, we present a model for syntactic category learning that builds on the merit of semantic bootstrapping and fully exploits the formal distributional properties of syntactic categories. The CFL model makes use of the Tolerance Principle to form category frames from lexical distributions. It is able to provide a highly accurate set of words and syntactic labels, from which additional distributions such as morphology can be extracted and integrated into the model.

Future research will explore the properties of the CFL model on other, typically more diverse, languages. We are currently exploring models that do not assume a pre-specified set of syntactic categories but postulate new categories on the basis of formal productivity. It is also important to bear in the mind that frames, while useful, are linear and cannot fully capture the distributional regularities in language. They are really stepping stones toward the development of grammar: We believe that the general framework of the CFL model can adapted for hierarchical structures and remain effective.

References

Anderson, Stephen R. (1969). *West Scandinavian vowel systems and the ordering of phonological rules*. Unpublished doctoral dissertation, MIT.

Brown, Roger. (1957). Linguistic determinism and the part of speech. *The Journal of Abnormal and Social Psychology*, *55*(1), 1–5.

Brown, Roger. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Carey, Susan E. (2009). *The origin of concepts*. New York: Oxford University Press.

Cartwright, Timothy A., & Brent, Michael R. (1996). Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition*, *63*(2), 121–170. doi: 10.1016/s0010-0277(96)00793-7

Chemla, Emmanuel, Mintz, Toben H., Bernal, Savita, & Cristophe, Anne. (2009). Categorizing words using 'frequent frames': what crosslinguistic analyses reveal about distributional acquisition strategies. *Developmental Science*, *12*(3), 396–406. doi: 10.1111/j.1467-7687.2009.00825.x

Chomsky, Noam. (1955). *The logical structure of linguistic theory.* (Ms., Harvard University and MIT. Revised version published by Plenum, New York, 1975)

Christodoulopoulos, Christos, Goldwater, Sharon, & Steedman, Mark. (2010, October). Two decades of unsupervised POS induction: How far have we come? In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 575–584). Cambridge, MA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D10-1056

Christodoulopoulos, Christos, Roth, Dan, & Fisher, Cynthia. (2016, August). An incremental model of syntactic bootstrapping. In *Proceedings of the 7th workshop on cognitive aspects of computational language learning* (pp. 38–43). Berlin: Association for

Computational Linguistics. Retrieved from https://aclanthology.org/W16-1906 doi: 10.18653/v1/W16-1906

Comrie, Bernard. (1989). *Language universals and linguistic typology*. Blackwell, Oxford.

Dye, Cristina, Kedar, Yarden, & Lust, Barbara. (2019). From lexical to functional categories: New foundations for the study of language development. *First Language*, *39*(1), 9–32.

Emond, Emeryse, & Shi, Rushen. (2021). Infants' rule generalization is governed by the Tolerance Principle. In Danielle Dionne & Lee-Ann Vidal Covas (Eds.), *Proceedings of the 45nd annual Boston University Conference on Language Development* (p. 191-204).

Gerken, LouAnn, Landau, Barbara, & Remez, Robert E. (1990). Function morphemes in young children's speech perception and production. *Developmental psychology*, *26*(2), 204.

Gerken, LouAnn, Wilson, Rachel, & Lewis, William. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, *32*(2), 249-268.

Gleitman, Lila R. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3-55. Retrieved from https://doi.org/10.1207/s15327817la0101_2 doi: 10.1207/s15327817la0101_2

Gómez, Rebecca L, & Lakusta, Laura. (2004). A first step in form-based category abstraction by 12-month-old infants. *Developmental science*, *7*(5), 567–580.

Haghighi, Aria, & Klein, Dan. (2006). Prototype-driven learning for sequence models. In *Proceedings of the human language technology conference of the north american chapter of the acl* (pp. 320–327).

Hall, Jessica, Van Horne, Amanda, & Farmer, Thomas. (2018). Distributional learning aids linguistic category formation in school-age children. *Journal of child language*, *45*(3), 717.

Harris, Zellig S. (1955). From phoneme to morpheme. *Language*, *31*(2), 190–222.

Köhne, Judith, Trueswell, John C, & Gleitman, Lila R. (2013). Multiple proposal memory in observational word learning. In *Proceedings of the 35th annual meeting of the cognitive science society. austin, tx: Cognitive science society.*

Koulaguina, Elena, & Shi, Rushen. (2019). Rule generalization from inconsistent input in early infancy. *Language Acquisition*, *26*(4), 416–435.

Macnamara, John. (1982). *Names for things: A study of human learning*. Cambridge: MIT Press.

MacWhinney, Brian. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Maratsos, Michael, & Chalkley, Mary Ann. (1980). The internal language of children's syntax: The nature and ontogenesis of syntactic categories. In Katherine Nelson (Ed.), *Children's language* (Vol. 2, p. 127-214). Cincinnati, OH: Gardner.

Marcus, Gary F, Vijayan, Sugumaran, Rao, S Bandi, & Vishton, Peter M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80.

Marcus, Mitchell P., Marcinkiewicz, Mary Ann, & Santorini, Beatrice. (1993, June). Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, *19*(2), 313–330.

Mintz, Toben H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117. doi: 10.1016/s0010-0277(03)00140-9

Mintz, Toben H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117.

Mintz, Toben H, Newport, Elissa L, & Bever, Thomas G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, *26*(4), 393–424.

Nivre, Joakim et al. (2017). Universal dependencies 2.1.

Pinker, Steven. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

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

Reeder, Patricia A, Newport, Elissa L, & Aslin, Richard N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive psychology*, *66*(1), 30–54.

Rondal, Jean A, & Cession, Anne. (1990). Input evidence regarding the semantic bootstrapping hypothesis. *Journal of Child Language*, *17*(3), 711–717.

Saffran, Jenny R., Aslin, Richard N., & Newport, Elissa. (1996). Statistical learning by by 8-month-old infants. *Science*, *274*, 1926-1928.

Santelmann, Lynn, & Jusczyk, Peter W. (1998). Sensitivity to discontinuous dependencies in language learners: Evidence for limitations in processing space. *Cognition*, *69*(1), 105-134.

Schuler, Kathryn D., Yang, Charles, & Newport, Elissa. (2016). Testing the tolerance principle: Children form productive rules when it is more computationally efficient to do so. In *Proceedings of the annual meeting of the cognitive science society.* Philadelphia, PA.

Shi, Rushen, Cutler, Anne, Werker, Janet, & Cruickshank, Marisa. (2006). Frequency and form as determinants of functor sensitivity in english-acquiring infants. *The Journal of the Acoustical Society of America*, *119*(6), EL61–EL67.

Shi, Rushen, & Melançon, Andréane. (2010). Syntactic categorization in French-learning infants. *Infancy*, *15*(5), 517–533.

Shi, Rushen, Werker, Janet F, & Cutler, Anne. (2006). Recognition and representation of function words in english-learning infants. *Infancy*, *10*(2), 187–198.

Soderstrom, Melanie, Blossom, Megan, Foygel, Irena, & Morgan, James L. (2008). Acoustical cues and grammatical units in speech to two preverbal infants. *Journal of Child Language*, *35*(4), 869-902.

Stevens, Jon Scott, Gleitman, Lila R, Trueswell, John C, & Yang, Charles. (2017). The pursuit of word meanings. *Cognitive science*, *41*, 638–676.

Stumper, Barbara, Bannard, Colin, Lieven, Elena, & Tomasello, Michael. (2011). Frequent frames in german child-directed speech: A limited cue to grammatical categories. *Cognitive science*, *35*(6), 1190–1205.

Theakston, Anna L., Lieven, Elena V. M., Pine, Julian M., & Rowland, Caroline F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language*, *28*(1), 127–152. doi: 10.1017/S0305000900004608

Van Heugten, Marieke, & Johnson, Elizabeth K. (2010). Linking infants' distributional learning abilities to natural language acquisition. *Journal of memory and language*, *63*(2), 197–209.

Yang, Charles. (2002). *Knowledge and learning in natural language*. Oxford: Oxford University Press.

Yang, Charles. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. Cambridge, MA: MIT Press.

Proceedings of the 46th annual Boston University Conference on Language Development

edited by Ying Gong
and Felix Kpogo

Cascadilla Press Somerville, MA 2022

ISSN 1080-692X
ISBN 978-1-57473-077-7 (2 volume set, paperback)

Ordering information

To order a copy of the proceedings or to place a standing order, contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, sales@cascadilla.com, www.cascadilla.com