An Inference-rules based Categorial Grammar Learner for Simulating Language Acquisition

Xuchen Yao* **Jianqiang Ma*** **Sergio Duarte*** **Çağrı Çöltekin**
University of Groningen
{x.yao,j.ma,s.r.duarte.torres}@student.rug.nl, c.coltekin@rug.nl

Abstract

We propose an unsupervised inference rules-based categorial grammar learning method, which aims to simulate language acquisition. The learner has been trained and tested on an artificial language fragment that contains both ambiguity and recursion. We demonstrate that the learner has 100% coverage with respect to the target grammar using a relatively small set of initial assumptions. We also show that our method is successful at two of the celebrated problems of language acquisition literature: learning English auxiliary fronting in polar interrogatives and English auxiliary word order.

1 Introduction

Grammatical inference is a field of machine learning that collects several methodologies and algorithms to learn formal languages. These techniques have been applied to learn language representations in several domains as biological sequence data (Park, 2001), robotics (Zender et al., 2008), and XML processing, among others. Besides practical engineering oriented applications of grammar inference in computational linguistics, a salient application is the explanation of learning natural language phenomena.

Children are able to acquire one or more languages to which they are exposed to without explicit instruction. Furthermore, *poverty of stimulus* argument (Chomsky, 1980), claims that the input received by children lacks critical information necessary for learning languages. Nativist theories account for these phenomena assuming that a large amount of language specific information is known innately by children. Empiricist theories, on the other hand, postulate general cognitive mechanisms as the basis of language acquisition.

As it is well known in machine learning, there is no single general purpose algorithm that can be applied to all possible problems (commonly referred to as *no free lunch theorem*). The learning process has to incorporate an initial *inductive bias*. The nature of inductive bias in the setting of language acquisition is not clear. Conceivably, the answer to this question should eventually come from neuroscience. However, as this does not seem to be possible in the near future, computational models of language can provide an insight into the nature of inductive bias needed for acquiring natural languages. Consistent computational models can indeed increase our understanding of the mechanism that humans employ to learn and use languages (Lappin and Shieber, 2007), particularly the mechanism involved in the acquisition of the first language.

Two of the popular test examples used in support of the argument of poverty of stimulus (APS) are the learning of auxiliary fronting in polar interrogatives and auxiliary word order. In this paper we show a computational learning method that deals with these two phenomena and is able to learn ambiguous and recursive artificial grammars. The method presented in this work is based on learning a categorial grammar.

Categorial Grammar (CG) is a lexicalized grammar formalism with a high level of transparency between syntax and semantics. These features make CG an attractive formalism for computational studies of language acquisition. The lexicalized nature of the CG reduces learning syntax to learning a lexicon, while the close connection between syntax and semantics helps learning one using the other.

One of the earliest studies of CG learners was proposed by Buszkowski and Penn (1989). Their

*Yao, Ma, and Duarte acknowledge financial support from the EU Life Long Learning Program through scholarships for the European Masters Program in Language and Communication Technologies (LCT).

Proceedings of the 18th Annual Belgian-Dutch Conference on Machine Learning, pages 29–37,
Tilburg, May 2009.

system used unification of type-schemes to determine categorial grammars from functor-argument structures. Kanazawa (1998) extended this algorithm and employed partial unification to learn from strings of words. A number of studies (e.g., Waldron (1999); Villavicencio (2002); Buttery (2006)) followed similar approaches to learn CG based grammars. Waldron (1999) used a rule-based method to infer a CG from input labeled with basic syntactic types. Villavicencio (2002) proposed a method that improves the performance of Waldron's system by describing an unconventional universal grammar based on CG, and using semantically annotated input. Watkinson and Manandhar (2000) presented an unsupervised stochastic learner which aims to learn a compact lexicon. They assumed that the set of possible categories are known, which maps the problem of grammar induction to categorization. The system achieved perfect accuracy in an artificial corpus. However, its performance dropped to 73.2% in lexicon accuracy and 28.5% in parsing accuracy when tested on the more realistic LLL corpus (Kanazawa, 1998).

We propose an unsupervised method to learn categorial grammars. The learner is provided with a set of positive sentences generated by a target grammar. Unknown categories are learned by applying a set of inference rules incrementally. When there are multiple choices, a *simple category preference* (SCP) principle that is inspired by the minimum description length (MDL) principle (Rissanen, 1989) is used to minimize the size of the grammar. We show that the learner is able to infer recursive and ambiguous grammars. Moreover, this method is capable of learning two well known linguistic phenomena: English auxiliary word order and English polar interrogatives. The method learns both phenomena successfully from a set of input sentences that are considered insufficient for these tasks.

The structure of this paper is as follows: Section 2 gives a short introduction to CG. Section 3 describes our learning architecture. Section 4 presents three experiments followed by a discussion of the results in Section 5. In the last section we provide brief conclusions and address future directions.

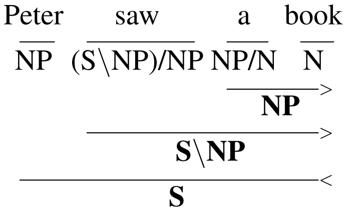


Figure 1: Example derivation for sentence *Peter saw a book.*

2 Categorial Grammar

Categorial grammar (Ajdukiewicz (1935; Bar-Hillel (1953)) is a lexicalized grammar formalism. CG describes all the language specific syntactic information inside the lexicon, leaving only a small number of universal rules outside the lexicon. We present a very brief introduction to CG here; more comprehensive descriptions can be found in Moortgat (2002) and Wood (1993).

Every word in a CG lexicon is assigned to a syntactic category. A limited set of categories constitutes the *basic categories* of the grammar. For example, **S** (sentence), **NP** (noun phrase), **N** (noun) are commonly assumed to be the basic categories for English. *Complex categories*, such as **NP/N**, **S\NP** and **(S\NP)\(S\NP)** are formed by combining any two CG categories with a forward (/) or backward (\) slash. Given the lexicon with categories of this form, the only rules of the CG are given in (1). An example derivation can be seen in Figure 1.

(1) *Function application rules*
Forward $A/B \quad B \quad \rightarrow \quad A \quad (>)$
Backward $B \quad A\backslash B \quad \rightarrow \quad A \quad (<)$

CG as described above is weakly equivalent to Context Free Grammars, and cannot model the complexity of natural languages adequately. However, CG is powerful enough for the linguistic phenomena presented here. It should be also noted that there are extensions such as *Combinatory Categorial Grammar* (CCG, (Steedman (2000); Steedman and Baldridge (2005)) that provide necessary descriptive and theoretical adequacy by introducing additional operations. In this work, we learn classical Categorial Grammars, while making use of some of the CCG operations, namely *composition*, *type raising*, and *substitution*, during the learning process. The additional CCG operations used are given in (2).

(2) a. *Function composition rules:*

$$\begin{array}{lllll} \text{Forward} & A/B \;\; B/C & \rightarrow & A/C & (>\mathbf{B}) \\ \text{Backward} & B\backslash C \;\; A\backslash B & \rightarrow & A\backslash C & (<\mathbf{B}) \end{array}$$

b. *Type raising rules:*

$$\begin{array}{lllll} \text{Forward} & A & \rightarrow & T/(T\backslash A) & (>\mathbf{T}) \\ \text{Backward} & A & \rightarrow & T\backslash(T/A) & (<\mathbf{T}) \end{array}$$

c. *Substitution rules:*

$$\begin{array}{lllll} \text{Forward} & (A/B)/C \;\; B/C & \rightarrow & A/C & (>\mathbf{S}) \\ \text{Backward} & B\backslash C \;\; (A\backslash B)\backslash C & \rightarrow & A\backslash C & (<\mathbf{S}) \end{array}$$

3 Learning by Inference Rules

The learning method presented in this paper is a rule-based unsupervised lexicalized grammar inference system. The input to the system is a set of grammatical sentences of the target language. The system learns a CG lexicon containing words assigned to (possibly multiple) CG categories. The set of lexical categories is not known in advance, it is generated by a number of inference rules which are the center of our learning algorithm.

In this section we first introduce a series of inference rules used to perform grammar induction. Then we will present the complete learning architecture along with an example demonstrating the learning process.

3.1 Grammar Induction by Inference Rules

Our inference rules work when there is only one unknown category in the input. Then a category for the unknown word is proposed by the inference rules. In the rule descriptions below, the letters *A, B, C and D* represent known categories, **X** represents the unknown category.

(3) *Level 0 inference rules:*

$$\begin{array}{llllll} B/A & \mathbf{X} & \rightarrow B & \Rightarrow & \mathbf{X} = A & if A \neq S \\ \mathbf{X} & B\backslash A & \rightarrow B & \Rightarrow & \mathbf{X} = A & if A \neq S \end{array}$$

(4) *Level 1 inference rules:*

$$\begin{array}{llllll} A & \mathbf{X} & \rightarrow B & \Rightarrow & \mathbf{X} = B\backslash A & if A \neq S \\ \mathbf{X} & A & \rightarrow B & \Rightarrow & \mathbf{X} = B/A & if A \neq S \end{array}$$

We define *level* as the number of *functioning* slash operators in a category. A slash operator *functions* if it takes an argument. Consequently, the basic categories are level 0. The category $S\backslash NP$ belongs to level 1. Note that the category of adverbs $(S\backslash_f NP)\backslash_f(S\backslash NP)$ belongs to level 2. Although it has three slashes, only the slashes marked with subscript $_f$ are functioning, i.e. can be used in a derivation.

Level 0 and level 1 inference rules can be successfully used to learn the category of intransitive verbs, such as *slept* in *Peter slept*. The condition $if \quad A \neq S$ in (3) and (4), prevents learning a large number of incorrect categories. For example, $S\backslash S$ for the word *well* from *Peter slept well*. As stated before, the category of adverbs belongs to level 2, so we need a level 2 inference rule to learn this category.

(5) a. *Level 2 side inference rules:*

$$\begin{array}{llllll} \mathbf{X} \;\; A \;\; B & \rightarrow & C & \Rightarrow & \mathbf{X} = (C/B)/A \\ A \;\; B \;\; \mathbf{X} & \rightarrow & C & \Rightarrow & \mathbf{X} = (C\backslash A)\backslash B \end{array}$$

b. *Level 2 middle inference rule:*

$$\begin{array}{llllll} A \;\; \mathbf{X} \;\; B & \rightarrow & C & \Rightarrow & \mathbf{X} = (C\backslash A)/B \end{array}$$

Level 2 inference rules are divided into two parts: the *side rule* and the *middle rule*, depending on whether an unknown category is at the beginning/end of a sentence or in the middle.

Notice that in (5b) the category $(C/B)\backslash A$ is as viable as the inferred category $(C\backslash A)/B$. In (5b), and (6b) presented later, we pick the *right-combining* rule.[1]

It might seem that using (5b) we can learn the category of $(S\backslash S)/NP$ for the preposition *with* from the sentence *Peter slept with Mary*. But this will not happen: the level 2 inference rule is implemented by recursively calling level 0 and level 1 inference rules, which all have the condition $if \quad A \neq S$ to prevent generating the category $S\backslash S$. As a matter of fact, none of the level 0-2 rules could help learning the category of *with* from the sentence *Peter slept with Mary*. So we need to use a level 3 inference rule.

(6) a. *Level 3 side inference rules:*

$$\begin{array}{llllll} \mathbf{X} \;\; A \;\; B \;\; C & \rightarrow & D & \Rightarrow & \mathbf{X} = ((D/C)/B)/A \\ A \;\; B \;\; C \;\; \mathbf{X} & \rightarrow & D & \Rightarrow & \mathbf{X} = ((D\backslash A)\backslash B)\backslash C \end{array}$$

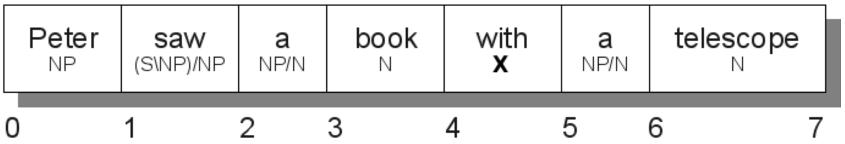
b. *Level 3 middle inference rules:*

$$\begin{array}{llllll} A \;\; \mathbf{X} \;\; B \;\; C & \rightarrow & D & \Rightarrow & \mathbf{X} = ((D\backslash A)/C)/B \\ A \;\; B \;\; \mathbf{X} \;\; C & \rightarrow & D & \Rightarrow & \mathbf{X} = ((D\backslash A)\backslash B)/C \end{array}$$

3.2 The Learning Architecture

The learning framework consists of three parts: *the edge generator*, *the recursive learner* and *the output selector*. A schematic description of the learning process is provided in Figure 2. Below we provide a detailed description of the three parts, along with demonstration of learning the ambiguous and recursive category of *with* from the input sentence given in Figure 3a.

[1] We realize that this is a rather arbitrary choice, and plan to relax it in future work. For now, we assume that this is a language specific choice learned before or during acquisition of phenomena presented here.



(a) Input string and index numbers. Note that the word *with*, marked as '**X**' is unknown at this point.

	span	rule used	category		span	rule used	category
1	(0, 1)	>T	S/(S\NP)	6	(0, 3)	>B	S/N
2	(0, 2)	>B	S/NP	7	(1, 4)	>	S\NP
3	(1, 3)	>B	(S\NP)/N	8	(2, 4)	<T	S/(S\NP)
4	(2, 4)	>	NP	9	(0, 4)	<	S
5	(5, 7)	>	NP	10	(0, 4)	>	S

(b) Generated edges in the chart.

	A		B		X		C	
	cat	span	cat	span	cat	span	cat	span
1	NP	(0, 1)	S\NP	(1, 4)	**((S\NP)\(S\NP))/NP**	(4,5)	NP	(5, 7)
2	S/(S\NP)	(0, 1)	S\NP	(1, 4)	((S\NP)\(S\NP))/NP	(4,5)	NP	(5, 7)
3	S/NP	(0, 2)	NP	(2, 4)	**(NP\NP)/NP**	(4,5)	NP	(5, 7)
4	S/NP	(0, 2)	S/(S\NP)	(2, 4)	(NP\(S/(S\NP)))/NP	(4,5)	NP	(5, 7)
5	S/N	(0, 3)	N	(3, 4)	**(N\N)/NP**	(4,5)	NP	(5, 7)

(c) Categories learned from the rule $A \quad B \quad \mathbf{X} \quad C \rightarrow D$ for the sentence in Figure 3a.

Figure 3: Example steps from the learning process.

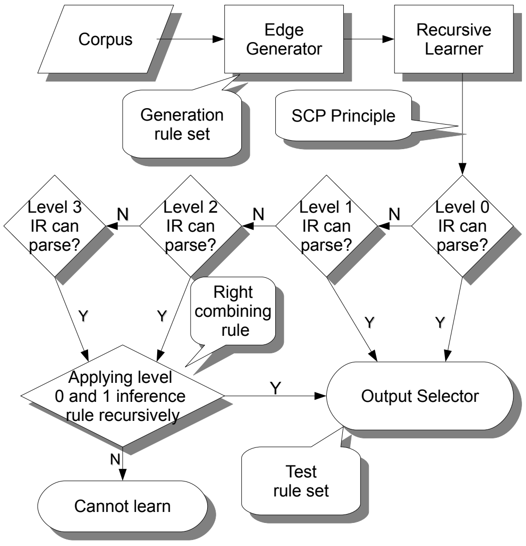


Figure 2: Learning process using inference rules.

The Edge Generator implements a variation of the CYK algorithm (Cocke (1969); Younger (1967); Kasami (1965)), which employs bottom-up chart parsing. Every known word in a sentence is an edge in the chart. The edge generator then tries to merge any consecutive edges into a single edge recursively. In order to produce as many edges as possible, besides function application rules (>,<), we have also used the composition (>B,<B), substitution (>S,<S) and the type raising (>T,<T) rules. Figure 3b shows all possible edges generated for the example in Figure 3a.

The Recursive Learner performs grammar induction by the rules given in Section 3.1. The learning process first tries to learn from level 0 or level 1 inference rules. If the unknown word cannot be learned by level 0 or level 1 inference rules, higher level rules are tried. Following *simple category preference* principle, if a category can be inferred with a lower level rule, we do not attempt to use higher level rules.

For the input in Figure 3a, the level 0 and level 1 inference rules are not enough. Only the level 3 middle inference rules (6b) can be applied. Figure 3c gives the list of all categories produced by

Peter	:= NP	*with*	:= (N\N)/NP
Mary	:= NP	*with*	:= (NP\NP)/NP
green	:= N/N	*with*	:= ((S\NP)\(S\NP))/NP
colorless	:= N/N	*sleep*	:= S\NP
book	:= N	*a*	:= NP/N
telescope	:= N	*give*	:= ((S\NP)/NP)/NP
the	:= NP/N	*saw*	:= (S\NP)/NP
run	:= S\NP	*read*	:= (S\NP)/NP
big	:= N/N	*furiously*	:= (S\NP)\(S\NP)

Table 1: Target grammar rules.

this inference rule.

The Output Selector tests the learned categories produced by the recursive learner and selects the ones that can be parsed using only function application rules. The categories that do not produce a valid parse with function application rules are discarded.

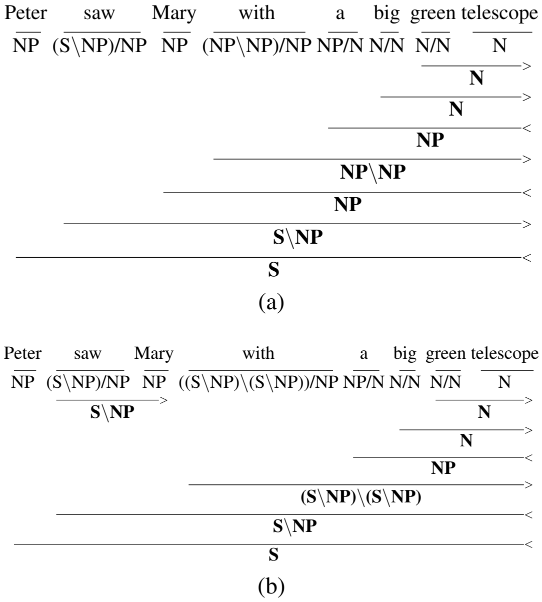
Not all rules in Table 3c are selected by output selector. We first remove duplicate categories generated in the previous phase. The category in row 2 is removed, as it is the same as the category in row 1. Furthermore, using only the function application rules, the sentence cannot be parsed with the category in row 4, so this category is discarded. Rows 1, 3 and 5 provide the learned categories.

4 Experiments and Results

We conducted three experiments with our learning system. In the first experiment, we tested the system's capabilities on an artificial language exhibiting a certain level of ambiguity and recursion. The second experiment tests model's ability to learn formation of English polar interrogative questions. In the third experiment, we tried to learn the English auxiliary order, another well known problem in language acquisition literature.

4.1 Experiment 1: Learning an Artificial Grammar

For this experiment, we have created a small English-like artificial grammar. The lexicalized grammar that is used as the target grammar for this experiment is listed in Table 1. The grammar includes both recursive rules and ambiguity. The input to the learner is generated by sampling 160 sentences that can be parsed using the target grammar. The input to the learner consist only of correct sentences. The input sentences are unlabeled, except for nouns (N) and proper names (NP). The learner is expected to converge to the target grammar.


Peter saw Mary with a big green telescope
NP (S\NP)/NP NP (NP\NP)/NP NP/N N/N N/N N
>
N
>
N
<
NP
>
NP\NP
<
NP
>
S\NP
<
S
Peter saw Mary with a big green telescope
NP (S\NP)/NP NP ((S\NP)\(S\NP))/NP NP/N N/N N/N N
>
S\NP
>
N
>
N
<
NP
>
(S\NP)\(S\NP)
<
S\NP
<
S

(a)

(b)

Figure 4: Two ambiguous parses of the sentence *Peter saw Mary with a big green telescope.*

After only a single pass through input sentences, all categories in our target grammar presented in Table 1 are learned correctly. The learned grammar is identical to the target grammar.

4.2 Experiment 2: Learning Auxiliary Verb Fronting

According to some theories of grammar, the English yes/no questions, such as (7b) are formed by *moving* the auxiliary verb in the corresponding declarative sentence (7a) to the beginning of the sentence. But when the subject comes with a relative clause, as in (7c), the auxiliary verb in the main clause, not the one in the relative clause, should be assigned the initial position (7d).

(7) a. *Peter is awake.*
b. *Is Peter awake?*
c. *Peter who is sleepy is awake.*
d. *Is Peter who is sleepy awake?*
e. **Is Peter who sleepy is awake?*

Starting with Chomsky (1965), it is frequently claimed that the input children receive during acquisition contains utterances of the form (7a), (7b) and (7c), but not (7d). Under these conditions, arguably, most plausible hypothesis to form the interrogative questions is fronting the first auxiliary. Even though this hypothesis generates sentences like (7e), Crain and Nakayama (1987) showed that children learn the more complicated ‘structure sensitive’ hypothesis. So, the nativist conclusion is that children must be making use of innate knowledge that language is *structure sensitive*. Together with many other APS claims, Pullum and Scholz (2002) have shown that the assumptions about the input were not right. Contrary to the claim, the relevant evidence, i.e. questions like (7d), was quite common in every corpus they have looked in.

In this study, we follow an approach similar to Clark and Eyraud (2006). We test our learner assuming that the APS claims are correct. We present our learner with the sentences of the form (7a), (7b) and (7c), and test if it can learn to correctly recognize and generate interrogatives with clauses. It should be noted that, our model is based on a *non-transformational* theory of grammar. Hence, the assumption that the question form is formed by transformations is not entertained. However, the learning problem is still valid: is it possible to learn how to form and interpret the interrogative questions without direct evidence provided in the data?

The experiment setting is the same as in experiment 1. The only additional information provided is the type of sentences, i.e. either given input is a declarative sentence (S) or a question (S_q). A fragment of the learned categories for the auxiliary verb *is* is given in (8). Figure 5 presents derivations of three sentences with the learned grammar. (a), (b) and (c) in Figure 5 are not surprising. However the correct parse in (d) shows that a simple learner can learn this type of constructions without being exposed to the same type of data. On the other hand, with all possible category assignments provided by learned lexicon, it is impossible to parse the incorrect sentences like (7e). Note that the learning is achieved by assigning CG categories the words in the input using the inference rules described in Section 3.1. The system learns a lexicalized grammar which is consistent with the inference rules.

Crucially, the learner does not make any explicit *structure dependence* assumption. Since the same categories can be assigned to the lexical items, or a collection of consecutive lexical items during derivation, the learned grammar can generate and recognize the correct forms without making any explicit assumptions about the sentence structure.

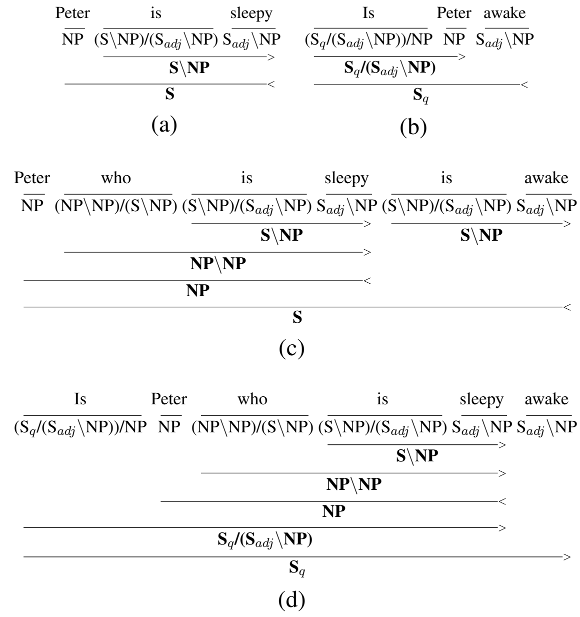

Peter is sleepy
NP (S\NP)/(S_adj\NP) S_adj\NP
S\NP
S
(a)
Is Peter awake
(S_q/(S_adj\NP))/NP NP S_adj\NP
S_q/(S_adj\NP)
S_q
(b)
Peter who is sleepy is awake
NP (NP\NP)/(S\NP) (S\NP)/(S_adj\NP) S_adj\NP (S\NP)/(S_adj\NP) S_adj\NP
S\NP
S\NP
NP\NP
NP
S
(c)
Is Peter who is sleepy awake
(S_q/(S_adj\NP))/NP NP (NP\NP)/(S\NP) (S\NP)/(S_adj\NP) S_adj\NP S_adj\NP
S\NP
NP\NP
NP
S_q/(S_adj\NP)
S_q
(d)

Figure 5: Derivations of the correct sentences in (7).

Assigning correct categories to the input words is enough for correctly producing and interpreting previously unseen sentence forms.

(8) *is* := $(S\backslash NP)/(S_{adj}\backslash NP)$
Is := $(S_q/(S_{adj}\backslash NP))/NP$

4.3 Experiment 3: Learning Correct Word Order

The difficulty of learning English auxiliary order has also been used as a support for argument of poverty of stimulus, and hence for linguistic nativism. Introduced first by Kimball (1973), the problem can be summarized as follows: the English auxiliary verbs *should*, *have* and *be* occur exactly in this order and all of them are optional. The claim is that while sentences containing a single auxiliary (9a–9c) or two auxiliaries (9d–9f) are present in the input, sequences of three auxiliaries (9g) are not ‘frequent enough’. Hence, it is not possible to learn the correct three-auxiliary sequence from the input alone.

(9) a. *I should go.*
b. *I have gone.*
c. *I am going.*
d. *I have been going.*
e. *I should have gone.*
f. *I should be going.*
g. *I should have been going.*
h. **I have should been going.*

should	$:= (S_s\backslash NP)/(S\backslash NP)$
should	$:= (S_s\backslash NP)/(S_h\backslash NP)$
should	$:= (S_s\backslash NP)/(S_b\backslash NP)$
have	$:= (S_h\backslash NP)/(S\backslash NP)$
have	$:= (S_h\backslash NP)/(S_b\backslash NP)$
be	$:= (S_b\backslash NP)/(S\backslash NP)$

Table 2: Categories of some auxiliary verbs.

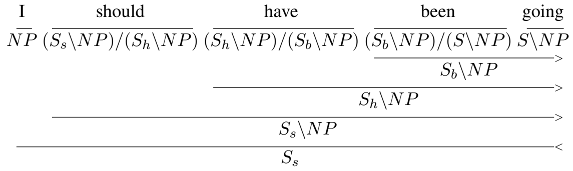

I
should
have
been
going
NP
$(S_s\backslash NP)/(S_h\backslash NP)$
$(S_h\backslash NP)/(S_b\backslash NP)$
$(S_b\backslash NP)/(S\backslash NP)$
$S\backslash NP$
>
$S_b\backslash NP$
>
$S_h\backslash NP$
>
$S_s\backslash NP$
<
S_s

Figure 6: Derivation of the correct word order.

The argument is controversial, and Pullum and Scholz (2002) have shown that there are more three-auxiliary sequences in the input than claimed. In this study, we choose another method to test the argument: we present our learner with sentences containing only one or two auxiliaries (as in (9a-9f)), and we test if it can correctly recognize and generate sentences with three auxiliaries. The setup is essentially the same as in experiment 2.

Table 2 presents a fragment of the learned grammar. The derivation of the sentence (9g) using the learned grammar is given in Figure 6. As can be verified easily, the lexicalized grammar presented in Table 2 would not allow sequences as in (9h). The categories assigned to auxiliary verbs by the learner completely, and correctly cover the English auxiliary order.[2]

Success of the learner is again due to its assignment of words to syntactic categories. The categories induced from one- and two-auxiliary sequences in a logical way extend naturally to the three-auxiliary sequences.

5 Discussion

We have presented a *simple*, *unsupervised* and *generalizable* method to learn natural language grammar. In this section, we will discuss these aspects of the system in the light of results presented in the previous section.

The model presented here is simpler than many alternatives. The CG rules used are based on a simple consistent notion of function application with sensitivity to spatio-temporal order. Learning proceeds by trying to classify the unknown items in the input by using a series of logical inference rules. The inference rules are simple and even intuitive. Though there is no solid psychological evidence that children learn a grammar in an inference-rule-based way, rule-based learning seems to be one of the methods in children's repository of learning methods starting as early as 7-months after birth (Marcus et al., 1999). We do not claim that rule-based learning is the only liable method, nor we claim that these rules are implemented in children's mind as we have done in a computer program. However, the 100% correct results in parsing and generation by our model suggests that it is sufficient to assume that children use a small set of rules together with a plausible inference procedure for learning the categories of unknown words. The only other additional piece in our learning algorithm is a preference towards simpler categories.

One characteristic of our method is that it is unsupervised. Raw text is fed to the learner without any extra syntactic structure, semantic annotation or negative evidence. The only additional cue we provide is marking the *basic* classes of nouns, and proper names. This can be justified on the grounds that nouns are learned earlier than other categories across the languages (Gentner, 1982). Arguably, children may have already learned these classes before acquiring the syntax of their language. Additional information, such as semantically annotated data is used in some of the related studies (e.g. Villavicencio (2002); Buttery (2006)). We have obtained good results from this unsupervised learning process. Knowing that input to the children is richer than raw text, if we give more hints of the learning material to the learner, our learner should get even better results in a more realistic learning environment.

In this study, we do not make use of any statistical information present in the input. As in

[2] An alternative approach would be assuming that the sequences like '*should have been*' are learned as single units, at least at the beginning of the learning process. Lexical items spanning multiple input units are considered by some of the related learners (e.g. Zettlemoyer and Collins (2005); Çöltekin and Bozsahin (2007)). However, to be compatible with the original claim, the learner presented in this paper assigns categories to single input units.

some of the related studies (e.g. Clark and Eyraud (2006)),[3] our non-statistical learner is adequate for the purposes of demonstrating the learnability of certain syntactic constructs. However, this results in two important limitations of our learner as a model of human language acquisition. First, it is known that humans make use of statistical regularities in the language for learning diverse linguistic tasks (Saffran, 2003; Thompson and Newport, 2007). Second, natural language data is noisy, and like any other rule-learner our algorithm is not able to deal with noise. Even though these look discouraging at first sight, it should be noted that it is easily extendible to incorporate statistical learning methods, for example, by learning a stochastic CG. We plan to improve the learning system in this direction, which, as well as allowing the learner to deal with noise in the input, would also allow us to lift or relax some of the assumptions we have made in this work. Others applied statistical methods to CG induction experiments successfully (Osborne, 1997; Watkinson and Manandhar, 2000).

In this paper we demonstrated that a fragment of the natural language syntax can be learned with a simple unsupervised learning method. This method performs well on learning two well-known examples of difficult to learn syntactic phenomena, namely the formation of English interrogative questions, and English auxiliary order. These phenomena are considered difficult to learn in the absence of critical data. Provided exactly with the type of data that was considered inadequate for the task, the learner was still able to learn the phenomena using only the simple mechanism described above. Even if APS claims are correct, children can still learn correct form with a simple inference mechanism.

6 Conclusion

We described a method to learn categorial grammars using inference rules of different levels according to the number of functional operators needed in the target category. Our method obtains a coverage of 100% on the target grammar. We use simple logical and intuitive inference rules to solve the problem of unknown categories in the input. The only additional aid provided to our learner is the simple category preference. Using only this set of initial assumptions, our system is also able to learn two phenomena that have been considered difficult to learn. As well as being able to recognize and generate English interrogative sentences with relative clauses without being exposed to the same type of sentences, the learner is also able to infer English auxiliary order correctly without being presented with all possible sequences. Our results demonstrate that a learner with simpler assumptions than the ones commonly assumed in the language acquisition literature is able to learn some of the difficult constructions found in natural language syntax. Putting aside the debate over the existence of ceartain type of evidence in the language acquisition process, our learner shows that exact experience is not always important for learning: some simple but logical inference rules are enough to help children deduce the correct syntactic structures.

However, it is necessary to note that our system has a number of limitations. First, these results were obtained using data that was generated artificially. Second, since we do not use any statistical rules, our system is not robust against noise. Using statistical patterns in the input language, it may also be possible to relax some of the assumptions presented here. These mark one direction in future work: developing the algorithm further to make use of statistical learning methods, and evaluating it on real data.

References

Kazimierz Ajdukiewicz. 1935. Die syntaktische Konnexität. *Studia Philosophica*, 1:1–27.

Yehoshua Bar-Hillel. 1953. A quasi-arithmetical notation for syntactic description. *Language*, 1:47–58.

Wojciech Buszkowski and Gerald Penn. 1989. Categorial grammars determined from linguistic data by unification. Technical report, Chicago, IL, USA.

Paula J. Buttery. 2006. Computational models for first language acquisition. Technical report, University of Cambridge, Churchill College.

Noam Chomsky. 1965. *Aspects of Theory of Syntax*. Cambridge, MA: MIT Press.

Noam Chomsky. 1980. *Rules and Representations*. Columbia University Press, New York, NY.

Alexander Clark and Rémi Eyraud. 2006. Learning auxiliary fronting with grammatical inference. In *Proceedings of CoNLL*, pages 125–132, New York.

[3] Unlike Clark and Eyraud (2006), our learner does not focus on learning a single phenomena but a complete grammar of the input language. The underlying grammar formalism also sets our study apart from theirs.

John Cocke. 1969. *Programming languages and their compilers: Preliminary notes*. Courant Institute of Mathematical Sciences, New York University.

Çağrı Çöltekin and Cem Bozsahin. 2007. Syllables, morphemes and Bayesian computational models of acquiring a word grammar. In *Proceedings of 29th Annual Meeting of Cognitive Science Society*, Nashville.

Stephen Crain and Mineharu Nakayama. 1987. Structure dependence in grammar formation. *Language*, 63(3):522—543.

Dedre Gentner. 1982. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. Kuczaj, editor, *Language development*, volume 2. Erlbaum, Hillsdale, NJ.

Makoto Kanazawa. 1998. *Learnable classes of categorial grammars*. Cambridge University Press, New York, NY, USA.

Tadao Kasami. 1965. An efficient recognition and syntax analysis algorithm for context-free languages. Technical Report AFCRL-65-758, Bedford, Massachusetts.

John P Kimball. 1973. *The Formal Theory of Grammar*. Englewood Cliffs, NJ: Prentice-Hall.

S. Lappin and S. Shieber. 2007. Machine learning theory and practice as a source of insight into universal grammar. *Journal of Linguistics*, 43(2):pp. 393–427.

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

Michael Moortgat, 2002. *Encyclopedia of Cognitive Science*, volume 1, chapter Categorial grammar and formal semantics, pages 435–447. London, Nature Publishing Group.

Miles Osborne. 1997. Minimisation, indifference and statistical language learning. In *In Workshop on Empirical Learning of Natural Language Processing Tasks, ECML'97*, pages 113–124.

Jong C. Park. 2001. Using combinatory categorial grammar to extract biomedical information. *IEEE Intelligent Systems*, 16(6):62–67.

Geoffrey K. Pullum and Barbara C. Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, 19:9–50.

Jorma Rissanen. 1989. *Stochastic Complexity in Statistical Inquiry Theory*. World Scientific Publishing Co., Inc., River Edge, NJ, USA.

J. R. Saffran. 2003. Statistical language learning: mechanisms and constraints. *Current Directions in Psychological Science*, pages 110–114, August.

Mark Steedman and Jason Baldridge. 2005. Combinatory categorial grammar. *Non-Transformational Syntax*.

Mark Steedman. 2000. *The syntactic process*. MIT Press, Cambridge, MA, USA.

Susan P. Thompson and Elissa L. Newport. 2007. Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, (3):1–42.

Aline Villavicencio. 2002. *The acquisition of a unification-based generalised categorial grammar*. Ph.D. thesis, University of Cambridge.

Ben Waldron. 1999. Learning grammar from corpora. Master's thesis, Cambridge University.

Stephen Watkinson and Suresh Manandhar. 2000. Unsupervised lexical learning with categorial grammars using the LLL corpus. In *Learning language in logic*, pages 218–233. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Mary McGee Wood. 1993. *Categorial Grammars*. Routledge, London.

Daniel H. Younger. 1967. Recognition and parsing of context-free languages in time $O(n^3)$. *Information and Control*, 10(2):189–208.

Hendrik Zender, Óscar Martínez Mozos, Patric Jensfelt, Geert-Jan M. Kruijff, and Wolfram Burgard. 2008. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems, Special Issue "From Sensors to Human Spatial Concepts"*, 56:62–67.

Luke S. Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the Twenty First Conference on Uncertainty in Artificial Intelligence (UAI-05)*.