BEYOND STATISTICAL LEARNING IN SYNTAX

ERI TAKAHASHI AND JEFFREY LIDZ

1. Introduction

Knowing the structural representation of sentences is a fundamental step for acquiring a language. However, the input to a child does not come with obvious labels to signal constituency–it seems like simple linear sequences of words. Since both words and constituent structure vary from language to language, children have to learn how words go together to form constituents in the particular language they are learning. Therefore, some learning mechanism must be present that guides the learner to build the correct phrase structure. What kind of input is necessary and what kind of information is used by children to come to the correct representation? It is most likely that children employ various kinds of information to arrive at the correct phrase structure representation–perhaps a combination of cues from prosody, function words, agreement morphology, semantics and distribution. This paper will focus on distributional cues to phrase structure.

Recent studies in artificial language learning have shown that distributional information can play a role in the acquisition of phonemes (Maye, Werker & Gerken 2002, Maye & Gerken 2000), word segmentation (Saffran, Aslin & Newport 1996), word categories (Mintz 2003) and syntax-like regularities (Gomez & Gerken 1999). In particular, it has been proposed that "transitional probabilities", which is a statistic that measures the predictiveness of the following element given a previous element, can be used by learners to successfully learn phrasal groupings of words (Thompson & Newport 2007) in miniature artificial languages.

Thompson & Newport (2007) showed that the adult subjects successfully learned the phrasal groupings of an artificial language based on the transitional probabilities. However, the artificial grammar in Thompson & Newport (2007) contained phrases with no internal structure and consequently leaves open the question of whether statistical cues to multiply embedded hierarchical structures can be detected by learners. So our question is: can learners use transitional probabilities to detect multiply embedded hierarchical structures?

A further question we have is: how is statistical information used by learners in the acquisition of phrase structure? Broadly speaking, we can think of at least two possibilities. One possibility is that each child has to discover the existence of phrase structure and its characteristics on the basis of distributional information alone ("phrase structure invention" hypothesis). A second possibility is that each child uses the input distribution to determine how the particular language maps words to structural descriptions of a highly restricted character ("phrase structure identification" hypothesis). These two hypotheses make distinct predictions about the nature of the mechanism for acquiring phrase structure representations. The "phrase structure invention" hypothesis predicts that the output of learning is determined solely by experience, while the "phrase structure identification" hypothesis holds that the output of learning derives from an interaction between the input and an experience-independent representational system. The current paper presents two new experiments to investigate (a) what kinds of distributional information can be used to identify hierarchical phrase structure and (b) whether learners use the distributional information to discover the existence of phrase structure or to map the experience onto a template.

2. Experiment 1

Experiment 1 explores whether the statistical cues to multiply embedded hierarchical structures can be detected by learners. Two miniature artificial languages–Grammar 1[1] and Grammar 2–were created. The two grammars share the identical word classes and lexical items, which were adapted from Thompson & Newport (2007). Each word class contained three nonsense lexical items.

Word Class	A	B	C	D	E	F
	KOF	HOX	JES	SOT	FAL	KER
	DAZ	NEB	REL	ZOR	TAF	NAV
	MER	LEV	TID	LUM	RUD	SIB

Table 1: Nonsense words assigned to each word class

The basic phrase structure trees for Grammar 1 and Grammar 2 are given below.

[1] We adapted the artificial languages from Morgan & Newport (1981), Morgan et al. (1987, 1989) and Saffran (2001).

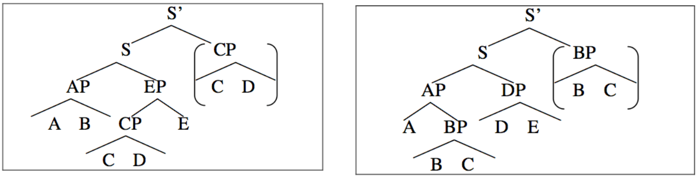

Fig 1: PS trees for Grammars 1 and 2

The canonical sentence in both grammars are identical–*ABCDE*. Grammars 1 and 2 differ only in constituent structure. For example, while AB is a constituent in Grammar 1, it is not in Grammar 2. Additionally, the grammars display nested hierarchical structure. In Grammar 1, a phrasal unit EP consists of an E word and another phrase CP, which in turn consists of C and D.

These grammars incorporate four types of manipulations which (a) made certain constituents optional, (b) allowed for the repetition of certain constituents, (c) substituted proforms for certain constituents and (d) moved certain constituents. For example in Grammar 1, the constituent AP can be replaced by a proform *ib*. As for the movement operation, the EP can be moved to the front in Grammar 1 and the DP can be moved in Grammar 2.

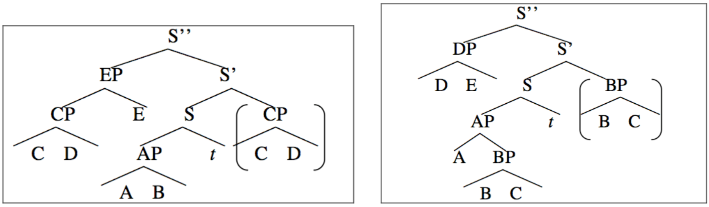

Fig 2: PS trees involving movement in Grammars 1 and 2

Eighty sentences from each language were picked as a presentation set. Incorporating all the manipulations discussed above resulted in the higher TPs between words within phrases compared with the TPs across phrases. Within a phrase, the TP is always 1.00 whereas TPs across phrase boundaries are substantially lower. The TP patterns of the presentation set are given below.

	A-B	B-C	C-D	D-E
Forward TP	1.00	0.24	1.00	0.25
Backward TP	1.00	0.19	1.00	0.34

Table 2: Transitional probabilities for 80 input sentences in Grammar 1

	A-B	B-C	C-D	D-E
Forward TP	0.33	1.00	0.15	1.00
Backward TP	0.18	1.00	0.16	1.00

Table 3: Transitional probabilities for 80 input sentences in Grammar 2

The sentences lacked any prosodic cues to phrase boundaries. The 80 sentences were randomized and repeated six times in a random order to form the familiarization input of approximately 36 min in duration.

Forty-four native English speakers participated in Experiment 1. Half of the participants were randomly assigned to hear Grammar 1 during the familiarization and the other half were assigned to Grammar 2. Both Grammar 1 and Grammar 2 subjects received the identical test items. The test was a forced-choice test. There were various test types: *Fragment test, Movement test* and *Substitution test*. In each test trial, participants heard a pair of word-sequences. Participants were instructed to choose the one they think belonged to the language they had just heard.

The Fragment Test was designed to assess the extent to which participants represented the input language in terms of phrasal groupings. Each trial consisted of two fragments, one that was a phrasal constituent in the input language and the other that was often a legal sequence but not a constituent in the input language.

	Grammatical in Grammar 1	Grammatical in Grammar 2
1	AB (e.g. *KOF HOX*)	BC (e.g. *NEB REL*)
2	CD (e.g. *JES SOT*)	DE (e.g. *SOT FAL*)

Table 4: Fragment test

A constituent in Grammar 1 (e.g. AB) is not a constituent in Grammar 2. Consequently, the correct answer for Grammar 1 was always the incorrect answer for the Grammar 2 condition, and vice versa for all test items.

To ensure that the performance on this test is a result of phrasal knowledge rather than frequency effects, we controlled the frequencies with which both groups of fragments appeared in the input. Specifically, *none* of the test items appeared in the input. Thus, the TP between any two neighboring words in all test items was 0. For example, the sequence of word categories AB may have appeared in the familiarization input, however, the sequence of the actual word tokens *KOF HOX* was never included in the input.

The Movement Test was designed to assess the extent to which participants allowed phrasal constituents to undergo a movement operation as opposed to non-constituents. This test was adapted from Morgan & Newport (1981) and Morgan et al. (1987, 1989). Each trial consisted of two sentences, one in which a constituent of the input language had been subjected to movement, and the other one in which a non-constituent of the input language had been subjected to movement. None of the test sentences occurred during familiarization.

	Grammatical in Grammar 1	**Grammatical in Grammar 2**
1	CDEAB (e.g. *JES SOT FAL KOF HOX*)	DEABC (e.g. *SOT FAL KOF HOX JES*)
2	FAB (e.g. *KER KOF HOX*)	DEF (e.g. *SOT FAL KER*)
3	CDEABCD (e.g. *JES SOT FAL KOF HOX JES SOT*)	DEABCBC (e.g. *SOT FAL KOF HOX JES HOX JES*)
4	FABCD (e.g. *KER KOF HOX JES SOT*)	DEFBC (e.g. *SOT FAL KER HOX JES*)

Table 5: Movement test

The Substitution Test was designed to assess the extent to which participants allowed phrasal constituents to be replaced by proforms *ib* and *et*. Each trial consisted of two sentences, one in which a constituent of the input language was substituted for by a proform, and the other in which a non-constituent of the input language was substituted by a proform. None of the test sentences occurred during familiarization.

	Grammatical in Grammar 1	**Grammatical in Grammar 2**
1	ib CDE (e.g. *ib JES SOT FAL*)	ABC ib (e.g. *KOF HOX JES ib*)
2	AB et E (e.g. *KOF HOX et FAL*)	A et DE (e.g. *KOF et SOT FAL*)
3	ib et E (e.g. *ib et FAL*)	A et ib (e.g. *KOF et ib*)

Table 6: Substitution test

If the subjects had learned the constituency, we predict that subjects in Grammar 1 would choose the correct answer for Grammar 1 as opposed to the answer for Grammar 2. Below, we report the percentage of times subjects chose the Grammar 1-compatible answers.

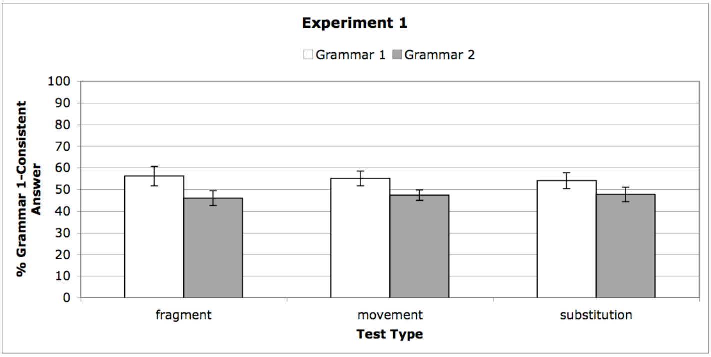

Fig 3: Results of Experiment 1

The participants who heard Grammar 1 as input chose the Grammar 1-consistent answers significantly more often than the participants who heard Grammar 2 during familiarization on the fragment test (*t*(1,42)=1.81, *p*<0.05) and the movement test (*t*(1,42)=1.84, *p*<0.05), but not on the substitution test (*t*(1,42)=1.30, *p*=0.10). Put another way, participants in both groups mostly chose the answers that were consistent with their input grammar significantly more often than chance.

Importantly, the results of experiment 1 indicate that participants acquired a grammar with nested hierarchical structure. The Fragment Test reveals that the Grammar 1 subjects learned that CD is a constituent. The Movement Test reveals that these subjects learned that CDE is a constituent. Putting these two conclusions together entails that subjects acquired nested hierarchical structures in which the constituent CD is a subpart of a larger constituent CDE.

The results of Experiment 1 suggest not only that can learners infer phrasal groupings on the basis of varying statistical pattern, but also that they can infer nested hierarchical structure. Furthermore, these results are not due to frequency effects, because none of the test items appeared in the input.

In Experiment 1, the Movement Test yielded a significant effect of learning. But one might argue that such result was due to the abundance of movement sentences in the input, and that subjects were simply choosing the ones that they were most familiar with. In fact, the presentation set in this experiment did include a large number of sentences that had undergone movement operation. In other words, the results of Experiment

1 are compatible with both the "phrase structure invention" hypothesis and the "phrase structure identification" hypothesis. We aim to tell apart these two possibilities in Experiment 2.

3. Experiment 2

In Experiment 2, we remove all the movement and substitution sentences from the input and examine whether subjects can nonetheless distinguish grammatical from ungrammatical movements and substitutions. There are two possible outcomes. The "phrase structure invention" hypothesis predicts that learners only allow new structures that have already been exhibited, since the learning is entirely based on the observed distribution. In specific, at test, participants should not be able to distinguish between the sentences in which a constituent or a non-constituent was manipulated, since both sentences are illicit because neither had appeared in the input. On the other hand, under a learning theory where statistics is just a path into innately known phrase structure system ("phrase structure identification" hypothesis), learners should allow new structures that have not yet been exhibited, as long as they are consistent with the learner's presuppositions about phrase structure representations. Thus, at test, participants are predicted to choose the sentences in which a constituent was moved over the sentences in which a non-constituent was moved. In this way, Experiment 2 examines whether statistical learning interacts with innate grammatical constraints.

The same artificial grammars, Grammar 1 and Grammar 2, were used. The only difference was that all examples generated via movement and substitution rules were excluded from the familiarization. Just like in Experiment 1, 80 sentences were picked as the presentation set. The resulting TP patterns of the presentation set are given below.

	A-B	B-C	C-D	D-E
Forward TP	1.00	0.28	1.00	0.24
Backward TP	1.00	0.24	1.00	1.00

Table 7: Transitional probabilities for 80 input sentences in Grammar 1

	A-B	B-C	C-D	D-E
Forward TP	1.00	1.00	0.22	1.00
Backward TP	0.22	1.00	0.24	1.00

Table 8: Transitional probabilities for 80 input sentences in Grammar 2

Forty-four native speakers of English participated in Experiment 2.

Half of the participants were randomly assigned to hear Grammar 1 during the familiarization and the other half were assigned to Grammar 2. The recording and the procedure for Experiment 2 were identical to those for Experiment 1. The test items were also identical to the ones in Experiment 1.

Participants in Grammar 1 chose the Grammar 1-consistent answers reliably more often than Grammar 2 participants on the movement test (*t*(1,42)=3.675, *p*<0.001), but not on the fragment test (*t*(1,42)=0.689, *p*=0.248) or the substitution test (*t*(1, 42)=-0.868, *p*=0.196).

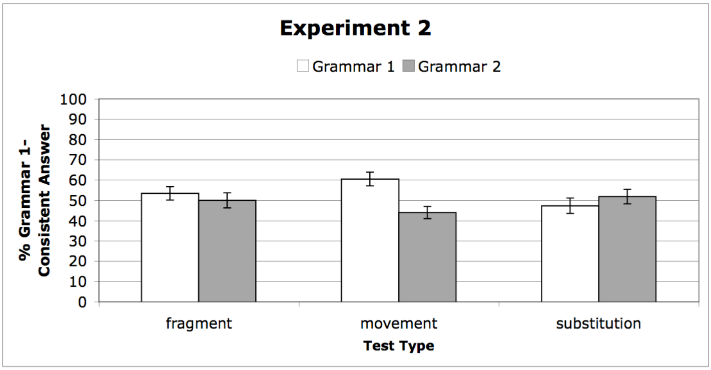

Fig 4: Results of Experiment 2

The participants in Experiment 2 were not given any movement rules or sentences in the input. Nevertheless, they chose the new sentences in which constituents, but not non-constituents, were moved. In addition, these results confirm that the success on the Movement Test in Experiment 1 was not due solely to the abundance of movement sentences in the familiarization.

The results of Experiment 2 are compatible only with the “phrase structure identification” hypothesis. Since learners did not see any input sentences involving movement, both test sentences–one that moved constituents and one that moved non-constituents–should be equally illicit under the first hypothesis. Thus, if statistical learning were purely input driven, participants should show no preference on the Movement test in Experiment 2. However, the performance was robust and well above chance. The information about constituency was contained in the input, but the information indicating that only constituents can move was not included in the input. In other words, the results of Experiment 2 are only consistent with the view that the syntactic inferences that learners make on the basis of distributional evidence go beyond simple statistical learning. Rather, these inferences appear to be driven by the learner's internally generated expectations about possible and impossible grammatical rules.

It is interesting that in the absence of movement and substitution rules in the input, participants were successful on the movement test but not on the substitution test. One possible reason for such asymmetry could be that, while you do not need input to infer that only constituents can be moved, you need sufficient information to infer that only constituents can be replaced by proforms. It could be that while you do not need any trigger to set off the constraints on movement, you need some kind of input to trigger the constraint on substitution to work. Another possibility is that learning pro-form substitution rules requires some identification of possible antecedents. In an artificial language experiment, no semantic information is provided that corresponds with the words or the sentences, but it might be that semantic information for identifying the referent of the proform is required.

4. Conclusion

The current paper asked whether it is possible to learn nested hierarchical structure on the basis of statistical distribution, whether learners make inferences about grammaticl structure that goes beyond what can be inferred just from the distributional evidence. The results of Experiment 1 showed that transitional probability can be a cue to not only the phrasal bracketing but also hierarchical constituent structure. Additionally, these results also show that it is possible to learn the structure of a sentence without relying on word meanings. The results of Experiment 2 showed that movement in the input is not required for the acquisition of constituent structure and to learn what constituents can undergo movement. It was also suggested that there is a contrast between movement-rule learning and substitution-rule learning. The results from Experiment 2 suggest that learners can project what they have learned based on the distributional information to novel structures they have not yet seen. But such projection to new structures occurred only within what is allowed by inherent constraints on the learner. This suggests that the statistical signature of phrase structure on surface strings serves as a cue for highly abstract knowledge that goes beyond what can be inferred from statistical evidence alone. In other words, statistical learning is an important component of phrase structure learning that works in tandem with the learner's inherent constraints on possible phrase structure representations.

The current findings suggest that transitional probability provides a useful cue to phrase structure. Furthermore, the present paper provides novel evidence that statistical learning interacts with innate constraints on phrase structure and movement rules, which suggests that distributional information is used as a path for selecting phrase structure representations from an inherently constrained hypothesis space.

References

Gomez, R. & Gerken, L.A. (1999). Artificial grammar learning by one-year-olds leads to specific and abstract knowledge. *Cognition, 70*, 109-135.

Maye, J. & Gerken, L.A. (2000). Learning phonemes without minimal pairs. In S. C. Howell, S. Fish & T. Keith-Lucas (Eds.), *Proceedings of the 24th Boston University Conference on Language Development*, 522–533. Somerville, MA: Cascadilla Press.

Maye, J., Werker, J. & Gerken, L.A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*, 101–111.

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

Morgan, J. L., Meier, R. P., & Newport, E. L. (1987). Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology, 19*, 498–550.

Morgan, J. L., Meier, R. P., & Newport, E. L. (1989). Facilitating the acquisition of syntax with cross-sentential cues to phrase structure. *Journal of Memory and Language*, *28,* 360–374.

Morgan, J. & Newport, E. L. (1981). The role of a constituent structure in the induction of an artificial language. *Journal of Verbal Learning and Verbal Behavior*, *20*, 67–85.

Saffran, J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language, 44*, 493–515.

Saffran, J., Aslin, R., & Newport, E. (1996a). Statistical learning by eight-month-old infants. *Science, 274*, 1926-1928.

Thompson, S. & Newport, E. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development, 3*, 1-42.