Using Child Utterances to Evaluate Syntax Acquisition Algorithms

Franklin Chang (chang.franklin@gmail.com)
NTT Communication Sciences Laboratories
2-4, Hikaridai, Seika-cho, Kyoto 619-0237, Japan

Elena Lieven (lieven@eva.mpg.de) and Michael Tomasello (tomas@eva.mpg.de)
Department of Developmental and Comparative Psychology, Max Planck Institute for Evolutionary Anthropology
Deutscher Platz 6, 04103 Leipzig, Germany

Abstract

Several algorithms for learning syntactic categories from distributional information were tested against utterances from adults and children in twelve typologically different languages. The evaluation measure that was developed allows one to examine word order constraints over a whole corpus and developmentally. By comparing several different algorithms of varying abstraction against actual corpora of children's speech, the evaluation measure determined that lexically specific knowledge is more advantageous than more broad-based category knowledge in predicting word order.

Introduction

There is a growing interest in unsupervised computational approaches to syntax acquisition (e.g., Mintz, 2003; Redington, Chater, & Finch, 1998). These systems collect statistical information from corpora and use that information to extract syntactic categories and constraints. These systems are evaluated by comparing their internal representations with syntactic representations that have been labeled or created by humans (e.g. tagged corpora). To compare these systems to human children, we would need to label child utterances with a set of syntactic representations. Since there is little agreement about the nature of syntactic representations in human children at each point in development (contrast Pinker, 1984, with Tomasello, 2003), it is difficult to use child utterances with computational approaches that are evaluated against human labeled representations.

One way to evaluate computational models that does not depend on tagged categories is suggested by connectionist approaches to syntax (Elman, 1990). Connectionist syntax acquisition models learn internal abstractions by making predictions over words. In these systems, the accuracy at predicting the order of words in a sentence is a measure of the system's language knowledge. One difficulty with scaling these systems up to real corpora is the use of neural assumptions such as gradual weight-based learning. In language models (e.g., n-grams), similar evaluation measures have been used, but since there are not enough constraints to predict the next word from the whole lexicon, researchers have typically looked at performance of these models with a subset of the data (e.g., frequent words). Here, we used prediction as the evaluation measure, but restrict it to the set of words from the utterance that we are trying to predict. This approach fits better with human performance and allows us to use corpora from multiple languages to test ideas from theories of syntax acquisition and adult sentence production.

Our version of this evaluation measure will be called Word Order Prediction Accuracy (WOPA) and it is based on models of sentence production. In production, one has a message that one wants to convey. To model the effect of the message in a rough way, we start with an unordered bag of words made up of words from the utterance that we want to predict, which we call the candidate set. Given this candidate set, the system has to try to predict the order of words in the sentence in a word-by-word fashion. After a word has been produced, it is removed from the candidate set, and the system tries to predict the next word in the sequence. After a sequence has been produced in this manner, we compare the sequence against the actual utterance in the corpus, and record whether it was correctly predicted or not. The Word Order Prediction Accuracy (WOPA) score is the number of correctly predicted utterances out of all the utterances of two words or more (one word utterances only have one ordering).

In this article, we will use WOPA to compare six syntax acquisition algorithms based on ideas from work in computational linguistics and child language. For input and testing, we will use utterances from child and parent interactions from twelve typologically-diverse languages. Our goal is to be able to determine which of these algorithms best matches the knowledge that may have yielded these corpora. More generally, we hope that this demonstration will suggest that WOPA can be useful as a general way of comparing computational approaches that use different internal representations.

Corpora

Computational learners should be tested against multiple typologically-different languages to avoid biases towards particular languages (e.g., English) or particular language typologies (e.g., languages without rich morphology). This is not usually done, because tagged corpora are not always available for different languages. WOPA does not evaluate against human-labeled syntactic tags and can be evaluated against raw word-separated corpora. Our corpora were twelve typologically diverse corpora

(Cantonese, Croatian, English, Estonian, French, German, Hebrew, Hungarian, Japanese, Sesotho, Tamil, Welsh) from CHILDES (Aldridge, Borsley, Clack, Creunant, & Jones, 1998; Berman, 1990; Demuth, 1992; Kovacevic, 2003; Lee, Wong, Leung, Man, Cheung, Szeto, & Wong, 1996; Miller, 1976; Miyata, 2000; Narasimhan, 1981; Réger, 1986; Suppes, Smith, & Leveillé, 1973; Theakston, Lieven, Pine, & Rowland, 2001; Vihman & Vija, in press). In addition, two larger English and German dense corpora from the Max Planck Institute for Evolutionary Anthropology were also used (Abbot-Smith & Behrens, in press; Lieven, Behrens, Speares, & Tomasello, 2003). These languages differ syntactically in important ways. German, Japanese, Croatian, Hungarian, and Tamil have more freedom in the placement of noun phrases (although the order is influenced by discourse factors) than English, French, and Cantonese. Several allow arguments to be omitted (e.g., Japanese, Cantonese). Several have rich morphological processes that lead to complex word forms (e.g. Croatian, Hungarian). Four common word orders are represented (SVO, SOV, VSO, no default order). Eleven genera are represented Chinese, Germanic, Finnic, Romance, Semitic, Ugric, Japanese, Slavic, Bantoid, Southern Dravidian, Celtic). All the corpora involved interactions between a target child and at least one adult that were collected from multiple recordings over several months or years. For each corpus, the *child* utterances were the target child utterances for that corpus, and the *adult* utterances were all other utterances.

Syntax Learners: Category-based algorithms

In this section, we will compare how well several different algorithms for learning distributional categories work in predicting the utterances in our fourteen corpora. WOPA evaluation depends on statistics collected from corpora. We will first describe the motivations for the kinds of statistics that will be collected. Second, we will describe the different algorithms for creating the categories that the statistics are collected on (Lexstat, Prevword, Freqframe, Token/Type, Type/Token). In addition, we will present a Chance learner that gives us a baseline level of performance. Then we will give an example of the statistics that are collected for a particular sentence, and also an example of how these statistics are used in production of a test sentence.

Our algorithms use statistics that are based on a dual-pathway model of sentence production (Chang, 2002; Chang, Dell, & Bock, 2006). In this *Dual-path* model, there are two pathways to guide word sequences. One pathway was called the sequencing system (using a simple recurrent network architecture, Elman, 1990), and this system learned how previous words constrained the next word in the sequences. The second pathway was called the meaning system, and it had a representation of the message that was to be produced. The meaning system was not able to encode sequencing information, and hence it set up a competition between all of the words that were activated by the message. So in this model, the sequencing system and meaning systems are independently trying to activate words, and the word with the best combined activation is selected as the next word in the utterance.

The Dual-path architecture was the basis for the statistics that were collected in our non-connectionist category-based statistical learners. To represent the sequencing pathway in the Dual-path model, a statistic was collected called the *context* statistic. This statistic represented how often the category of a word directly followed another word (akin to bigram statistics in computational linguistics). The other statistic represented the message pathway in the Dual-path model and it was called the *access* statistic. Unlike the context statistic, which only encodes the relationship between adjacent words, the access statistic encodes how often a category precedes other words in the sentence separated by any number of words (these statistics can encode long distance dependencies). The most important difference in the context and access statistics will be in how they are used in production, which we will describe after we describe the categorization algorithms and present an example of the statistics that they collect.

Lexstat Learner: A simple categorization algorithm for words is one that simply treats all words as separate categories (each category has one member). The Lexstat learner exemplifies this strategy by collecting statistics using the words themselves as the category. For example, the word “ate” would be a member of the *ate* category.

Prevword Learner: Many computational linguistic approaches make use of the preceding word as context for classifying the next word (Redington, Chater, & Finch, 1998). In our version of this learner, each word is categorized based on the most frequent previous word. If the most frequent previous word is “you”, then “ate” would be classified as a member of the YOU category.

Freqframe Learner: Mintz (2003) proposed that a frame made up of one word before and one word after was a better way to classify words into categories like nouns and verbs. For our version of this learner, each word was classified by the most frequent frame that surrounded it.

Token/Type Learner and Type/Token Learner: The slots in the Freqframe algorithm differ in their lexical diversity. Some slots seem to have a wide range of members, and some have a relatively small set. One way to measure lexical diversity is with the ratio of unique word types to the number of token words. So it might be useful to pick frames based on their lexical diversity of their slots. However, it is not yet clear whether we should prefer slots with a large lexical diversity or a small lexical diversity. A large lexical diversity may be evidence of a

general categorizer, or it might simply be a frame which accepts many different categories. Likewise, a frame with a small lexical diversity might be a selective categorizer, but it might also be a frame whose behavior is dominated by a few idiosyncratic members. To test both of these possibilities, two learners were created, one which categorizes words with the frame that has the highest lexical diversity (Type/Token Learner) and one which uses the frame with the lowest diversity (Token/Type Learner).

Now that the five categorization algorithms have been described, we can examine how the context and access statistics would be collected for these different learners. To make this concrete, let us work through how the statistics for "ate" would be calculated in the sentence "We ate the cake" (Table 1).

Table 1: Example of statistics incremented for word "ate". Capitalized Words are categories.

Learner	Context Statistic	Access Statistics
Lexstat	we -> ate	ate > the
		ate > cake
Prevword	we -> YOU	YOU > the
		YOU > cake
Freqframe	we -> YOU_IT	YOU _IT > the
		YOU _IT > cake
Token/Type	we -> SHE_IT	SHE_IT > the
		SHE_IT > cake
Type/Token	we -> YOU_IT	YOU _IT > the
		YOU _IT > cake

In the Lexstat Learner, the "we" to "ate" context statistic and the "ate" before "the" and the "ate" before "cake" access statistic would be incremented. In the Prevword approach, if the most frequent word preceding "ate" in the corpus is "you", then "ate" would be in the YOU category and the same statistics would be collected except that the YOU category would replace the "ate" category. In the Freqframe approach, if the most frequent frame that "ate" occurs in is "you ate it", then "ate" would be classified as an YOU_IT category, and corresponding statistics would be collected. The categories in the Type/Token and Token/Type Learners would depend on the number of members in each frame. Lets say that "ate" occurs also between "she" and "it", and the SHE_IT category has a frequency of 4 and 2 unique members and the YOU_IT has a frequency of 9 and 9 unique members. The Type/Token Learner would classify "ate" with YOU_IT (9/9 > 2/4), while the Token/Type Learner would classify with SHE_IT (4/2 > 9/9).

Since context and access statistics are just counts of how often the word and a category appear together in a particular order, the counts can vary greatly due to the frequency of two elements involved. To equate for this, we divide the context and access statistics by a count of how often two elements both occur in the same utterance. This helps to make the ordering statistics for low frequency elements equivalent to the ordering statistics for high frequency elements.

Before describing how these statistics are used, we should first introduce the last learner, which is the Chance learner. The Chance learner just estimates the probability of getting the sentence right by randomly generating an order. For example, a two word utterances has only two orders and therefore a 50% chance of getting the order right by guessing. A three word utterances has six orders and 16.7% chance of getting the right order. Chance sentence accuracy is then simply a function of the length of the utterance (% correct = 100/n!). If a word occurs more than once in an utterance, it is give a unique label (e.g., "the-1 boy saw the girl"), but either "the" and "the-1" are consider correct when "the" is expected. Hence the Chance learner is a bit lower than the actual chance level.

The creation of the categories and the collection of the statistics encompass the creation of the learner. Next, we need to test these learners and evaluate the results using WOPA. To give an example of this, we will work through the utterance "Do you want to throw something in the rubbish?" (Brian 3;5), which was correctly predicted by the Lexstat learner and which has never been produced by the adults in the corpus. Initially, the algorithm starts with the candidate set ("do", "in", "rubbish", "something", "the", "throw", "to", "want", "you") and uses the punctuation as the first previous word (question mark in this case). Since sentences tend to start with words that are related to their purpose (e.g., English questions tend to start with question words or verbs, while statements tend to start with pronouns or determiners), using the punctuation as the first previous word allows us to captures this relationship. Each of the candidate words has a choice value, and the word with the highest choice value is selected at each point in a sentence. The choice function for each word in the candidate set is incremented by the context statistic from the previous word. For example, English input to children has many questions that start with "do", so the context statistic from "?" to "do" will be strong, and that will increase the choice value for "do" at the beginning of this specific novel utterance. The choice function for each word in the candidate set is also augmented by the access statistics between this word and all the other candidate words. For example, "want" and "throw" both appear after pronouns like "you" in the input to the child. But when an utterance has both "want" and "throw", "want" tends to precede "throw" (as in the mother's utterance "I didn't **want** you to **throw** the string."). If "want" precedes "to", "throw", "something", "in", "the", and "rubbish" more than "throw" precedes "to", "want", "something", "in", "the", and "rubbish", then the access statistics will increase the value of the choice function for "want". Since the choice value of a particular word is changed by access statistics from all the words in the candidate set, the context statistic from the previous word is multiplied

by the length of the candidate set before being added to the choice value.

Our goal here is to see which of these proposed theories of category formation are best able to learn constraints that are implicit in the utterances in our corpora. The first test will be a situation which we called self-prediction where the input and the output for the algorithm was the same. This allows us to see to what extent our particular learners can account for the data under ideal input conditions. Self-prediction gives us a view of how consistent a corpus is with itself. For example, if a corpus has only two sentences: “it is here” and “here it is”, then the statistics in these two sentences are not going to be able to predict whether “here” goes before or after “it is”. Hence, self-prediction with this corpus will be at 50% (one of these two sentences will be incorrectly predicted). Notice that this is higher than chance (100/3! =16.7%), since the order for the words "it" and "is" is consistent with this corpus. If the orders in the corpora are consistent for each set of words and predictable with our learners, then we should expect that the self-prediction accuracy would be higher than the Chance learner.

To test whether the models differ from each other, t-tests were performed treating our fourteen corpora as the population. These t-tests tell us how likely we would see a difference between these learners if we selected a random corpus from the same population. By using typologically-different languages as our population, differences between our learners will generalize to other languages that come from that population.

Fig 1 shows that our learners were able to predict the order of words in these typologically-different languages. Using a learner that categorizes words using the previous word (Prevword) yields a 28% improvement over what would expected by chance ($t(13) = 11.37$, $p < 0.0001$). Restricting the categories with the following word (Freqframe) increases the prediction accuracy by 10% ($t(13) = 7.38$, $p < 0.0001$). Both of these algorithms pick the most frequent categorizers, but if we divide by the number of unique words in that frame and pick the frame with the best ratio (Token/Type), we get a further 8% improvement ($t(13) = 12,19$, $p < 0.0001$). The Token/Type Learner prefers frames which have few members, but are highly frequent. But an even higher improvement (5%) is reached if we pick frames with a high Type/Token ratio ($t(13) = 8.0$, $p < 0.0001$). Finally, a learner that just uses lexical-lexical statistics has the highest accuracy (71%) over all the category-based learners (7% higher than the Type/Token learner, $t(13) = 12.24$, $p < 0.0001$).

Another way to evaluate these algorithms is by examining how well the adult input can be used to predict the child’s output (Fig. 2). This is a stricter test, because there are words and utterances in the child’s speech which

Fig. 1: Adult-Adult Prediction for Category-based Learners

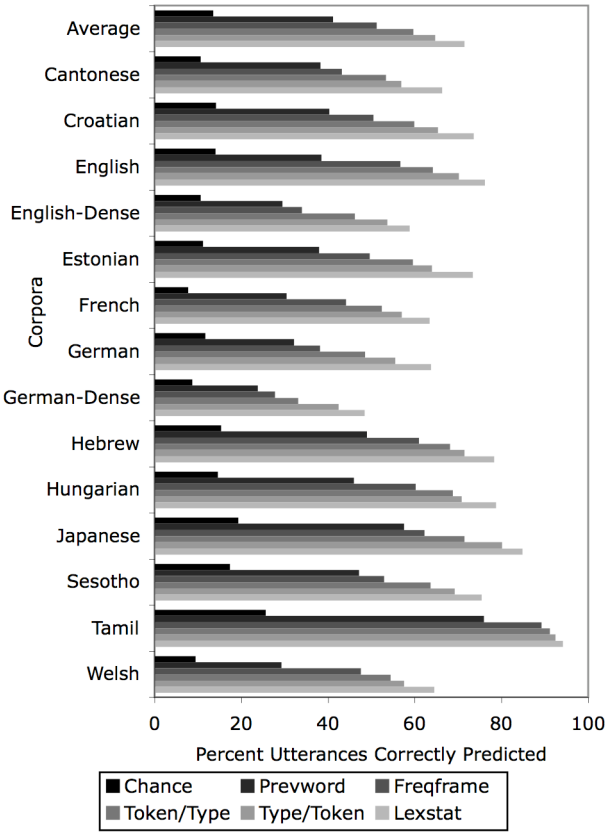
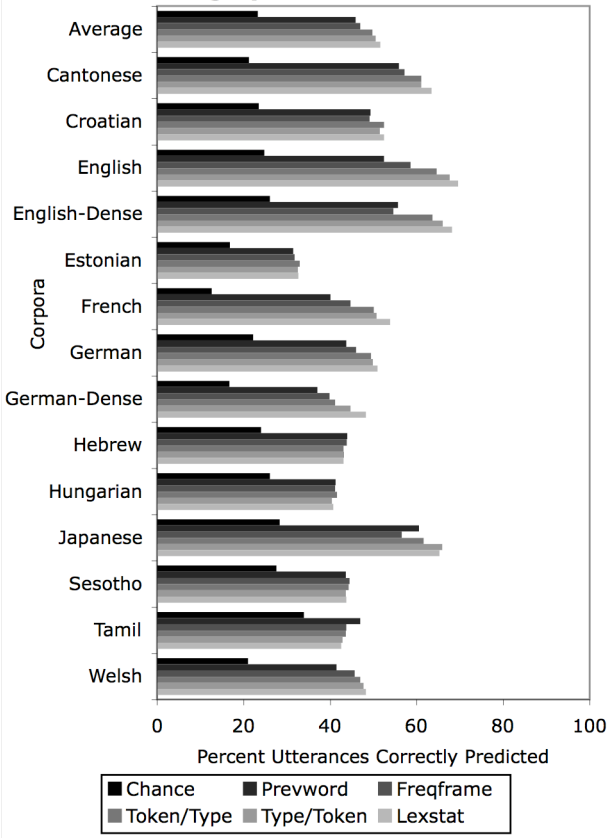


Fig. 2: Adult-Child Prediction for Category-based Learners



are not present in the adult speech, and hence this tests the systems' ability to generalize to novel sentences.

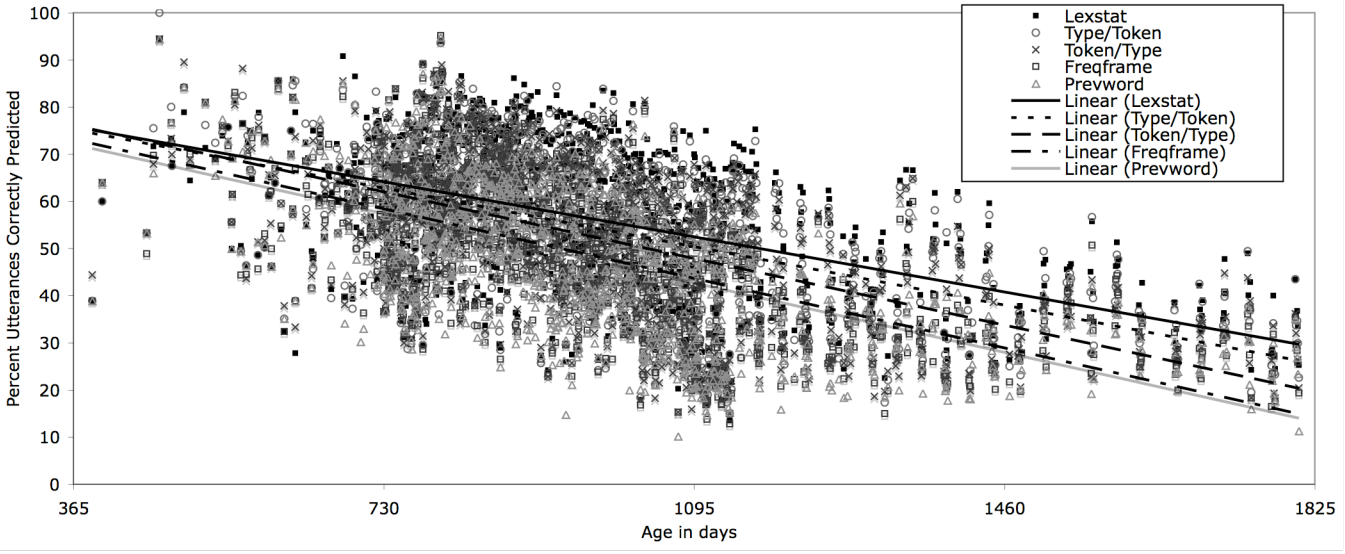
The average adult-child results suggest that rank order of learners is the same, even when tested on child utterances. But some of the adult differences between learners are no longer significant. Prevword and Freqframe Learners are the same ($t(13) = 1.31$, $p = 0.21$). Token/Type Learner is still better than the Freqframe Learner ($t(13) = 3.65$, $p = 0.003$). The Type/Token Learner is no different than the Token/Type Learner ($t(13) = 1.72$, $p = 0.11$). And the Lexstat Learner remains the better than the Type/Token Learner ($t(13) = 3.03$, $p = 0.01$). So it would seem that taking lexically-specific information into account, either in the ratio or in the statistics, is what yields improvement in the learners.

Another way to compare learners to see which learner best matches the child's syntactic development. To examine this, we calculated the prediction accuracy of all five learners for each day in each corpus (Fig. 3). Each of the corpora is collected at different frequencies (daily, monthly) and at different periods in each child's life between 1 and 5 years. For each learner, we estimate a linear regression that attempts to predict the accuracy level given the age of the child in days (computed with the age in days approximation: years * 365 + months * 31 + days). The slope of the regression tells us how consistent the algorithm is at predicting the utterances over development. Since the utterances that children produce are becoming more syntactically complex over time, the slope is typically negative, since the systems can usually predict simpler utterances better than longer and more complex utterances. The slope is independent from overall accuracy, because it is possible to have an algorithm which has a high overall accuracy and a low slope, and vice versa. To compare slopes, we use a t-test with accuracy on each day over all the corpora as the sample. This analysis tells us whether we would see a difference in these slopes if we were to sample another child from this age range and from a language with features that are similar to our typologically-diverse sample.

The Prevword and Frequent-frame learners had the same slope ($t(2344) = 0.03$, $p = .97$). And although the intercept for the Token/Type is higher than the Frequent-frame, the slope is the same ($t(2344) = 0.80$, $p = 0.43$). The Type/Token Learner has a more positive slope than the Token/Type Learner ($t(2344) = 2.0$, $p = 0.045$) and was not different from the Lexstat Learner ($t(2344) = 0.82$, $p = 0.41$). This suggests that the Type/Token and Lexstat learners are best able to account for the more complex utterances later in development, where the accuracy results for five algorithms diverge.

So in general, what do these results says about the match between these algorithms and child data? It is clear that category and statistics in the Prevword and Freqframe learners are useful in characterizing the orders in child speech, but at the same time, the tendency of these algorithms to discover broad categories (e.g., like nouns and verbs) makes it hard to order members of the same category relative to one another. The Token/Type and Frequent Frame algorithms both depend on high token frequency, but by dividing by the number of unique types, the Token/Type learner is able to yield more specific categories, which seems to increase the match with what children are producing. The Type/Token is a quite different algorithm from the Token/Type, as it prefers frames with a high lexical diversity in the slot. This preference actually makes the Type/Token learner more like the Lexstat learner. This is because frames with a single word member (like lexical items in the Lexstat Learner) have a high type/token ratio and therefore are sometimes selected by the Type/Token learner. Given the results here, we can say that we are better able to characterize the order of words in child and adult speech using more specific categories like words rather than with broad categories. So although none of the learners has

Fig. 3: Comparison of Developmental Trends in Syntactic Learners.



learned standard linguistic syntactic categories (e.g., nouns, verbs, adjectives, determiners), the ones that were closest to having these broad categories (e.g., Prevword, Freqframe) were less good at producing word order than those with non-standard categories (e.g., Lexstat, Type/Token). It maybe the case that combinations of broad and specific categories might work better, but more work is needed to specify how this is done.

Conclusion

Learning to order words is a crucial behavior in language acquisition and constrained word order is one important indicator of internal syntactic constraints. But instead of using word order for evaluation, most computational systems used abstract categories and structures for evaluation and these measures depend on theoretical considerations for their validity. Our approach takes advantage of the syntactic constraints on word order and therefore does not require human-labeled categories or structures for evaluation. In this work, we have demonstrated that WOPA evaluation measures can be used to compare six different learners with child and adult utterances in twelve typologically different languages. This evaluation measure provides several advantages. Instead of optimizing computational linguistics systems for the limited set of languages that are typically studied, we can use WOPA to compare systems against typologically different languages, allowing us to work towards a universal account of syntax acquisition. In addition, because WOPA is compatible between connectionist and non-connectionist approaches to language, it provides a way to combine and integrate these computational approaches. Finally, WOPA works on child utterances, which are rarely tested by computational systems, because they are difficult to tag with syntactic representations. Since children are the only known systems that can learn the syntax of any human language, it seems wise to use their utterances to help evaluate syntax acquisition algorithms.

References

Abbot-Smith, K., & Behrens, H. (in press). How known constructions influence the acquisition of new constructions: The german periphrastic passive and future constructions. Cognitive Science.

Aldridge, M., Borsley, R. D., Clack, S., Creunant, G., & Jones, B. M. (1998). The acquisition of noun phrases in Welsh. In Language acquisition: Knowledge representation and processing. Proceedings of GALA '97. Edinburgh: University of Edinburgh Press.

Berman, R. A. (1990). Acquiring an (S)VO language: Subjectless sentences in children's Hebrew. Linguistics, 28, 1135-1166.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. Cognitive Science, 26(5), 609-651.

Chang, F., Dell, G. S., & Bock, J. K. (2006). Becoming syntactic. Psychological Review, 113(2), 234-272.

Demuth, K. (1992). Acquisition of Sesotho. In D. Slobin (Ed.), The Cross-Linguistic Study of Language Acquisition (Vol. 3, pp. 557-638). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14(2), 179-211.

Kovacevic, M. (2003). Acquisition of Croatian in crosslinguistic perspective. Zagreb.

Lee, T. H. T., Wong, C. H., Leung, S., Man, P., Cheung, A., Szeto, K., et al. (1996). The development of grammatical competence in Cantonese-speaking children. Hong Kong: Dept. of English, Chinese University of Hong Kong. (Report of a project funded by RGC earmarked grant ,1991-1994).

Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. Journal of Child Language, 30(2), 333-367.

Miller, M. (1976). Zur Logik der frühkindlichen Sprachentwicklung: Empirische Untersuchungen und Theoriediskussion. Stuttgart: Klett.

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

Miyata, S. (2000). The TAI corpus: Longitudinal speech data of a Japanese boy aged 1;5.20 - 3;1.1. Bulletin of Shukutoku Junior College, 39, 77-85.

Narasimhan, R. (1981). Modeling language behavior. Berlin: Springer.

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

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. Cognitive Science, 22(4), 425-469.

Réger, Z. (1986). The functions of imitation in child language. Applied Psycholinguistics. , 7(4), 323-352.

Suppes, P., Smith, R., & Leveillé, M. (1973). The French syntax of a child's noun phrases. Archives de Psychologie, 42, 207–269.

Theakston, A., Lieven, E., Pine, J., & Rowland, C. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. Journal of Child Language, 28(1), 127-152.

Tomasello, M. (2003). Constructing a language: A usage-based theory of language acquisition. Cambridge, MA: Harvard University Press.

Vihman, M. M., & Vija, M. (in press). The acquisition of verbal inflection in Estonian: Two case studies. In N. Gagarina & I. Gluzow (Eds.), Verb Grammar in the Early Stages of Language Acquisition (Vol. 1-22). Dordrecht: Kluwer.