Towards a quantitative corpus-based evaluation measure for syntactic theories

Franklin Chang, Elena Lieven, and Michael Tomasello
Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Abstract

A corpus-based learner, inspired by a psycholinguistic model of sentence production, is able to predict the order of words in a large proportion of adult and child utterances in 5 typologically-different languages without using syntactic abstractions. Contrary to poverty of the stimulus arguments, learning on a small amount of adult input yields strong improvements in predicting the children's utterances. This system provides a way to evaluate learnability of word order constraints in typologically-different languages.

- **Question**: How are corpora used to study language development?
- **Traditional Response**: Compare the output of a syntactic learner against the categories in tagged corpora (Redinton, Chater, & Finch, 1998; Mintz, 2003).
- **Problem 1**: Categories in children might not match adult tags.
- **Problem 2**: Tagging is a subjective process performed by different people using different tag categories for each corpora (both within and across languages).
- How do we evaluate a syntactic learner without using tagged categories for comparison?

An Alternative Evaluation Measure: Word Order Prediction

Syntactic constraints are embodied in the order of words (e.g. determiners before nouns in English, verb in final position in Japanese). Word order prediction can be a way to evaluate syntactic constraints cross-linguistically. Connectionist models of syntax acquisition do word order prediction (Elman, 1990; Chang, 2002).

Sentence Production with Corpora: The Lexical Producer

Lexical producer is a corpus-based syntactic learner that is inspired by a psycholinguistically-motivated connectionist model of sentence production (Dual-path model, Chang 2002).

1. Production begins with a message: Unordered candidate list of words from the actual utterance.
2. For each candidate list word, the system applies two biases:
    1. Context bias: Given the previous word, how likely is this candidate word the next word (similar to bigrams). Akin to constraints in Dual-path model's sequencing system.
    2. Access bias: Does this candidate word tend to precede the other candidate words. Akin to competition during lexical access in Dual-path model's message system.
3. Sum together biases for each candidate word, and choose the most highest sum as the next word.
4. Remove actual word from candidate list, go to 2 and repeat.

Example of production in the Lexical Producer:

Target Utterance: "but i'm a big boy" (Thomas, 3 years, 11 months)

Candidate set	Prediction	Actual	Correct
• a, big, boy, but, i'm	but	but	√
• a, big, boy, i'm	I'm	I'm	√
• a, big, boy	a	a	√
• big, boy	boy	big	X
• boy	boy	boy	√

Lexical Producer: but I'm a boy boy = Incorrect
Random: I'm a big boy boy = Incorrect

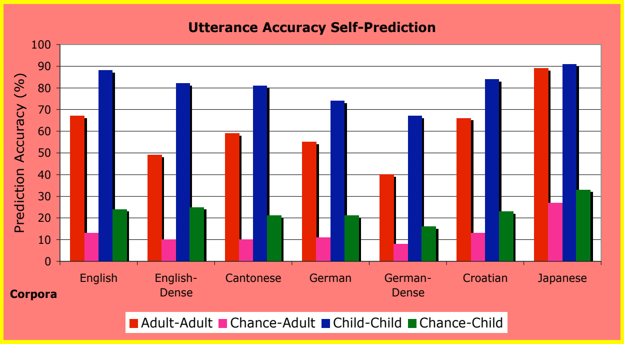
Syntactic learners should work in typologically-different languages.

Corpora	Syntax features
English (Theakston, et al., 2001)	fixed word order
English-Dense (Lieven et al., 2003)	"
Cantonese (Lee et al., 1996)	fixed word order, prodrop
German (Miller, 1976)	free word order
German-Dense (MPI-EVA)	"
Crotian (Kovacevic, 2003)	free word order, rich morphology
Japanese (Miyata, 1992, 1995)	free word order, prodrop

How consistent are children and adults in their ordering of words?

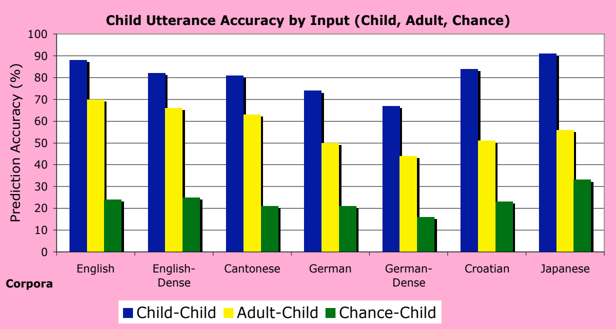
Use same corpus for training and test (self-prediction).

Finding: Children use the same word order for a set of words, 81% of the time. Adults use the same order 61% of the time.



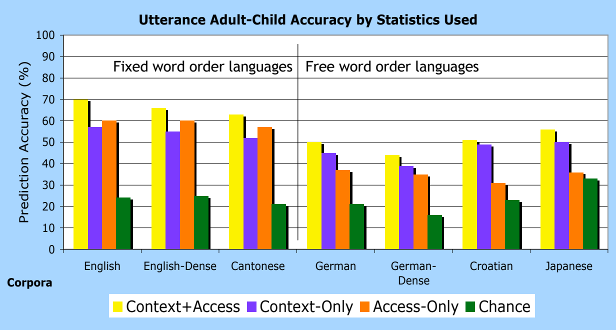
How helpful is the adult input for learning to order a child's utterances?

Finding: A small sample of adult data increases prediction accuracy for the child's utterances 34% over chance. More than half of the distance between self-prediction and chance for all corpora.



Do languages differ in how they use the statistics?

Finding: Context+Access is better than Context or Access alone. Fixed word order languages use Access more than Context, while free word order languages use Context more than Access.



Main Findings

- More than half of the word orders in corpora can be explained by lexically-specific ordering patterns without syntactic abstractions.
- Input may be impoverished with respect to syntactic abstractions, but not in terms of lexically-specific ordering regularities.
- Lexical producer can work equally well for fixed/free word order, prodrop, and rich morphology languages.

References

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

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

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

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

Presented at the 18th Annual CUNY Sentence Processing Conference (2005), Tucson, Arizona