Comparing different approaches for using n-grams in syntax acquisition

Franklin Chang (NTT Communication Sciences Laboratories, NTT Corp., Kyoto, Japan)

Abstract

•Although n-grams are important in both developmental psycholinguistic theories and computational linguistic approaches for syntax acquisition (Langkilde-Geary, 2002; Reali & Christiansen, 2005), there has been relatively little work comparing different n-gram algorithms in typologically-different languages using spoken utterances of the sort that children hear during language acquisition.
•Here, ten n-gram models were tested in 12 typologically-different languages within a sentence prediction evaluation task called the BIG-SPA task. We found that combinations of n-grams yielded better performance than individual n-grams, but these improvements were restricted to analytic languages. These results suggest that children learning synthetic languages may require very different mechanisms to acquire the word order of their language.

BIG-SPA syntax evaluation task

We compared various n-gram based learners using the BIG-SPA syntax evaluation task (Chang, Lieven, & Tomasello, in press).
Outline of BIG-SPA task:

1. Collect statistics from a corpus (e.g. n-grams).
2. Use statistics to generate an utterance incrementally from the set of word in the sentence (Bag-of-words Incremental Generation, or BIG).
3. Compare the generated sentence with the target sentence (Sentence Prediction Accuracy, or SPA):
   - If the sentence is an exact match, then the algorithm has knowledge that is sufficient to explain the word ordering knowledge that generated the sentence.

Why use the BIG-SPA task?

- Doesn't require labeled corpora (not theory dependent).
- Oriented towards syntax (grammaticality depends on the order of all of the words in an utterance).

Corpora

- Standard corpora in computational linguistics (e.g., Penn Treebank, Brown) do not resemble the input to human children.
- Used spoken adult-child CHILDES corpora from 12 typologically different languages
  - English, German, Cantonese, French, Japanese, Welsh, Croatian, Sesotho, Hebrew, Hungarian, Tamil
  - Two dense corpora from the MPI-EVA were also used.

Typologically Different Languages

- Word order variation
  - Limited (English, Cantonese) vs flexible (Tamil, Hungarian)
- Argument omission
  - All arguments can be omitted (Cantonese, Japanese), all arguments required (English, German).
- Morphology
  - Rich opaque morphology (Croatian, Hungarian), limited morphology (English, Cantonese)

N-gram Learners

Collected n-grams counts from adult utterances.
$C(w_{n-k} \ldots w_n)$ = frequency of n-gram for k=0,1,2,3,4,5
nwords = number of word tokens

N-gram sentence production: Pick the word in the bag-of-words with the highest value according the statistic in the learner. For 2-gram to 5-gram learners, the MLE equation was used.
k-gram = $C(w_{n-k} \ldots w_n)/C(w_n)$
Unigram learners used this equation: 1-gram = $C(w_n)$/ntotal

In computational linguisitics, n-grams are typically smoothed by combining higher order n-grams with lower order n-grams. Higher order n-grams are more specific, but lower n-grams are more likely to cover the words in the test set. Smoothed n-gram learners were simple combinations of the individual n-grams (e.g., 2+3+4-gram = 2-gram + 3-gram + 4-gram).

Summary of n-gram learners tested:

- Individual Statistic: 2-gram, 3-gram, 4-gram, 5-gram
- Combined Statistics: 2+3-gram, 2+3+4-gram, 2+3+4+5-gram
- Combined Statistics with unigram frequency: 1+2+3-gram, 1+2+3+4-gram, 1+2+3+4+5-gram

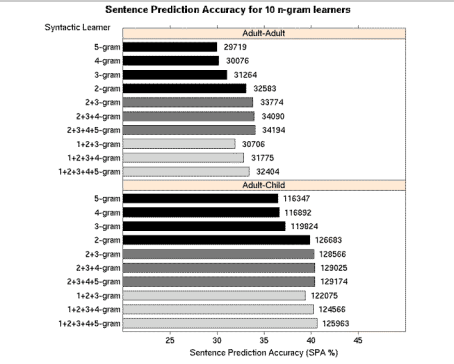
Testing

Two testing situations were used.

- Adult-Child test asked whether we can prediction the child's utterances from the statistics in their adult input.
- Adult-Adult test allowed us to see how well the system predicted 10% of the adult utterances from the rest of the adult input.
- Paired t-tests were used to compare different n-gram algorithms for each corpus.

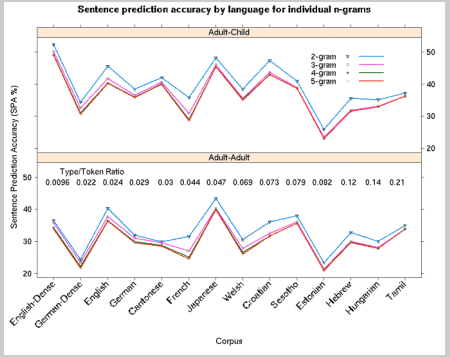
Results:

- Bigrams are much better than other individual n-grams (black bars below) because they are more likely to overlap between input and test utterances (e.g., 2-gram vs. 3-gram, $t(13) = 6.5$, $p < 0.001$)
- Smoothed algorithms (e.g., 2+3-gram, in dark-gray bars) are better than individual n-grams (e.g., 2-gram), because the higher order n-grams are more accurate than the bigrams (2+3-gram vs. 2-gram, $t(13) = 4.4$, $p < 0.001$).
- Unigrams seem to reduce the accuracy of smoothed n-grams (but not significantly, e.g. 1+2+3-gram vs. 2+3-gram, $t(13) = 1.6$, $p = 0.13$; light-gray bars).

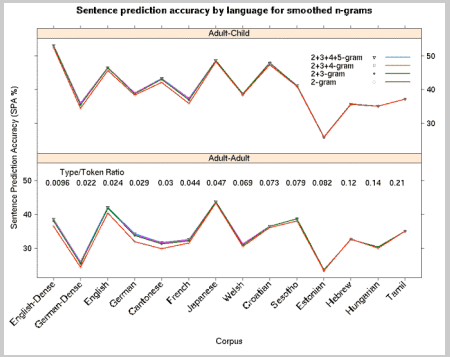


Corpus Comparison

To explore how different n-grams are used in different languages, we separated our results by language. Below are the individual n-grams results for Adult-Adult test and Adult-Child test. Trigrams seem to be more useful than higher order n-grams in the languages on the left side of the graph.



Also, we see a similar language difference for the smoothed algorithms below, where 2-grams and higher order smoothed algorithms are similar in the languages on the right, but more separated for the languages on the left.

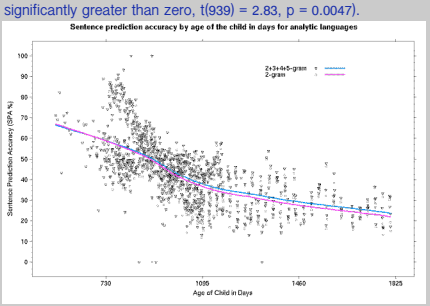


The languages on the left side are more *analytic* languages, which use separate function words to mark syntactic distinctions, while the languages on the right are more *synthetic*, using opaque morphology to mark the same distinctions. The ordering of languages from left to right is based on the ratio of word types to word tokens, where analytic languages tend to have low ratios relative to synthetic languages (the type/token ratios are shown in the figures).

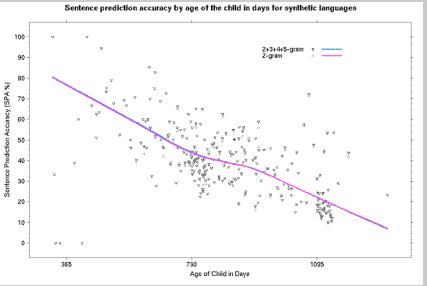
To test whether the algorithms differ depending on the analytic/synthetic distinction, we compared the difference between 2+3-gram and the 2-gram in synthetic and analytic languages. The difference between these algorithms was larger in the analytic languages (1.2) vs. the synthetic algorithms (0.25) ($t(8) = 4.4$, $p = 0.002$). N-grams seem to modulate the learning of word order in analytic languages, but do not seem to modulate the learning of synthetic languages.

Does the analytic/synthetic distinction matter for development?

We computed the BIG-SPA score for each day in each of the corpora for the 2-gram and the 2+3+4+5-gram learner. In the analytic languages, the utterances that children produced can be predicted better when higher order n-grams are used (slope of the difference between 2+3+4+5-gram and 2-gram over age is significantly greater than zero, $t(939) = 2.83$, $p = 0.0047$).



However, there is no difference between these algorithms for the synthetic languages (slope of difference, $t(252) = -0.58$, $p = 0.56$). Higher order n-grams are more rare in synthetic languages, since there are fewer function words and more morphologically opaque words.



Conclusions

- N-grams can be useful for predicting word order in typologically different languages.
- Smoothed n-grams are better than individual n-grams, but most of this benefit occurs for analytic languages.
- Higher order n-grams are not very useful in synthetic languages. These results suggest that children might require other statistical mechanisms to acquire these languages. See Chang, Lieven, & Tomasello (in press) for an example of a more powerful learner.

References

•Chang, F., Lieven, E., & Tomasello, M. (in press). Automatic evaluation of syntactic learners in typologically-different languages. Cognitive Systems Research.
•Langkilde-Geary, I. (2002). *An empirical verification of coverage and correctness for a general-purpose sentence generator.* Proceedings of the International Natural Language Generation Conference, New York City, NY.
•Reali, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science, 29*, 1007-1028.