Statistics vs. UG in language acquisition:
Does a bigram analysis predict auxiliary inversion?

Xuân-Nga Cao Kam

PhD Program in Linguistics

The Graduate Center,
City University of New York

xkam@gc.cuny.edu

Iglika Stoyneshka

PhD Program in Linguistics

The Graduate Center,
City University of New York

idst_r@yahoo.com

Lidiya Tornyova

PhD Program in Linguistics

The Graduate Center,
City University of New York

ltornyova@gc.cuny.edu

William Gregory Sakas

PhD Programs in Computer Science and Linguistics
The Graduate Center
Department of Computer Science,
Hunter College,
City University of New York

sakas@hunter.cuny.edu

Janet Dean Fodor

PhD Program in Linguistics
The Graduate Center
City University of New York

jfodor@gc.cuny.edu

Extended Abstract

Reali & Christiansen (2003, 2004) have challenged Chomsky's most famous "poverty of stimulus" claim (Chomsky, 1980) by showing that a statistical learner which tracks transitional probabilities between adjacent words (bigrams) can correctly differentiate grammatical and ungrammatical auxiliary inversion in questions like (1) and (2):

(1)   Is the little boy who is crying hurt?
(2)   *Is the little boy who crying is hurt?

No examples like (1) occurred in the corpus that R&C employed, yet the grammatical form was chosen by the bigram model in 92% of the test sentence pairs. R&C conclude that no innate knowledge is necessary to guide child learners in making this discrimination, because the input evidently contains enough <u>indirect</u> statistical information (from other sentence types) to lead learners to the correct generalization.

R&C's data are impressive, but there is reason to doubt that they extend to other natural languages or even to other constructions in English. While replicating R&C's Experiment 1 (see Data [A]), we discovered that its success rests on 'accidental' English facts.

Six bigrams differ between the grammatical and ungrammatical versions of a sentence. (The 6 relevant bigrams for the test sentence pair (1)/(2) are shown in Table 1.) However, 86% of the correctly predicted test sentences were definitively selected by the single bigram "who is" (or "that is"), because it occurred in the corpus and none of the remaining 5 bigrams did.

Distinctive bigrams in (1)	**who is**	is crying	crying hurt
Distinctive bigrams in (2)	who crying	crying is	is hurt

Table 1. Six bigrams that differentiate *Is the little boy who is crying hurt?* from *Is the little boy who crying is hurt?* The first sentence is selected (as grammatical) solely due to the high probability of *who is*.

It can be anticipated that when there is no bigram "who/that is" in the grammatical test

sentence (e.g., in relative clauses with object-gaps, auxiliaries such as *was*, *can*, or *do-support*), the learning will be less successful. Our results confirm this prediction: object relatives like (4) and (5), where "who/that is" is not present, were poorly discriminated (see Data [B]).

(4)   Is the wagon your sister is pushing red?
(5)   *Is the wagon your sister pushing is red?

Results for sentences with only main verbs, requiring do-support in question-formation, like (6) and (7), were also very weak (see Data [C]).

(6)   Does the boy who plays the drum want a cookie?
(7)   *Does the boy who play the drum wants a cookie?

Furthermore, the powerful effect of "who/that is" in R&C's experiment reflects no knowledge of relative clauses. It rests on the homophony of English relative pronouns with interrogative "who" and deictic "that". In R&C's training-set, the phonological/orthographic form "who" occurred as relative pronoun only 3 times, but as interrogative pronoun 44 times. R&C's analysis didn't differentiate these. (Similarly for "that": 14 relative versus 778 deictic or complementizer.)

In some languages relative pronouns are homophonous with other parts of speech (e.g., with determiners in German). We explored the possible effects of this by replacing the relative pronouns in the English corpus with "the". Discrimination between grammatical and ungrammatical aux-inversion was poor (see Data [D]).

Many human languages lack any such superficial overlaps with relative pronouns. So unless there are other cues instead, learning can be expected to be unsuccessful in those languages too. We tested this hypothesis in two ways:

(i) We distinguished relative pronouns from their non-relative homophones in English by coding the former as "who-rel" and "that-rel" in both the corpus and the test sentences. We found a greatly reduced ability to select the grammatical aux-inversion construction (see Data [E]).

(ii) We tested verb fronting in Dutch questions, using a Dutch corpus comparable to the English corpus used by R&C (the Groningen Dutch corpus from CHILDES; approximately 21,000 utterances of child-directed speech, age 1;8 to 1;11). Due largely to verb-final word order in relative clauses, there was no one distinctive bigram that could be relied on to predict the correct choice. Performance on a set of 20 items tested so far was no better than chance (see Data [F]). Clearly, the Dutch examples provided no alternative cues for selecting the grammatical version.

Thus, the success rate in R&C's experiment has very limited applicability. In general, bigram probability (or sentence cross-entropy, as computed in these experiments) is a poor predictor of grammaticality; e.g., the measure that prefers (1) over (2) **mis**-prefers (8) over (9):

(8)   *Scared you want to the doggie.
(9)   She can hear what we're saying.

We conclude that the bigram evidence against the poverty of the stimulus for language acquisition has not been substantiated to date. It remains to be seen whether richer statistics-based inductive models will offer more robust cross-language learnability.

References

Chomsky, N. (1980). in M. Piattelli-Palmarini, (1980) *Language and Learning: The Debate Between Jean Piaget and Noam Chomsky.* Cambridge: Harvard University Press.

Reali, F. & Christiansen, M. H. (2003). Reappraising Poverty of Stimulus Argument: A Corpus Analysis Approach. *BUCLD 28 Proceedings Supplement.*

Reali, F. & Christiansen, M. H. (2004). Structure Dependence in Language Acquisition: Uncovering the Statistical Richness of the Stimulus. *Proceedings of the 26th Annual Meeting of the Cognitive Science Society.*

Data

	% correct	% incorrect	% can't choose	# of sentence pairs tested to date	Experiment
A	87	13	0	100	Replication of R&C
B	33	15	52	100	Object-gap
C	50	50	0	50	Do-support
D	17	41	42	100	"The" replacement
E	17	39	44	100	Who-rel/That-rel
F	45	50	5	20	Dutch