Á. Gallego & D. Ott (eds.). 2015. *50 Years Later: Reflections on Chomsky's* Aspects. Cambridge, MA: MITWPL.
Available for purchase at http://mitwpl.mit.edu

INPUT AND ITS STRUCTURAL DESCRIPTION

CHARLES YANG
ALLISON ELLMAN
JULIE ANNE LEGATE
University of Pennsylvania

Does input data matter in language acquisition? Of course it does: a child born in Kansas learns English and a child born in Beijing learns Mandarin. But input data is hardly strings of words. According to the generative tradition of language acquisition outlined in *Aspects*,

> [T]he child approaches the data with the presumption that they are drawn from a language of a certain antecedently well-defined type, his problem being to determine which of the (humanly) possible languages is that of the community in which he is placed. (Chomsky, 1965:27)

That is, the primary linguistic data receives structural descriptions, which are determined by the strong generative capacity of the grammatical system under consideration (*Aspects*, §6 and §9).

While much attention has been devoted to the argument from the poverty of the stimulus (Chomsky, 1975), where the child demonstrates the mastery of linguistic knowledge in the absence of sufficient evidence (Legate and Yang, 2002), equally revealing observations can be made about the nature of the grammatical system when the child appears oblivious to an abundance of evidence in the input data. During the much studied null subject stage, for instance, English learning children systematically (and probabilistically) omit the use of the grammatical subject until around the age of 3 (Valian, 1991) even though adult speech almost always contains the subject. Children's behavior would be puzzling if they interpreted the linguistic data as surface patterns, especially since we know that children are exceptionally good at matching the probabilistic aspects of language use (Roberts and Labov, 1995, Smith et al., 2009). If, on the other hand, children analyze the input under the guidance of "the (humanly) possible languages" (including *pro*-drop and topic drop grammars, Jaeggli and Safir 1989), a string such as *They play soccer* and indeed almost all the English data would not differentiate among these options (Hyams, 1986) and the prolonged stage of subject omission is accounted for (Yang, 2002).

In this paper we present a case study that highlights the input data's explanatory limits when interpreted superficially, and explanatory power when given appropriate structural descriptions.

Our study concerns the acquisition of tense marking, and in particular past tense marking, in African American English (AAE) in contrast with Mainstream American English (MAE). Young children acquiring both varieties of English display a pattern, known as Root Infinitives (RI), whereby they produce a notable proportion of bare root verbs when a finite form is required. We show that the developmental patterns across dialects cannot be attributed to the statistical properties of specific linguistic forms but must make reference to an overarching and abstract system of grammar.

1 Input and Infinitives

The phenomenon of root infinitives (RI) has been and continues to be a major topic in language acquisition; see Phillips (1995) and Legate and Yang (2007) for review. For many languages in which the root verb needs to be tense marked, young children produce a significant number of verbs in the root clause that are nonfinite.

(1) a. Papa have it. (English)

b. thee drinken. (Dutch)
tea drink.INF

c. Dormir petit bébé. (French)
sleep.INF little baby

d. mein Kakao hinstelln. (German)
my chocolate.milk put.down.INF

e. Malon lauf. (Hebrew)
balloon fly.INF

An important aspect of the RI phenomenon is its gradient distribution, both within and across languages. Within a language, the use of RI is probabilistic rather than categorical, hence its alternative designation: "optional infinitive". The nonfinite verbal forms are generally used in co-existence with finite forms, and children gradually, rather than suddenly, exit from the RI stage as they grow older. Furthermore, the cross-linguistic distribution of RI is also continuous: some languages have much longer RI stage than others: for instance, the RI stage may persist in English and Dutch for over four years while the stage is only very briefly present in languages such as Italian and Spanish.

The variational model of language acquisition (Yang 2002, 2004) can incorporate the statistical properties of the primary linguistic data within an abstract grammar model, and explains the gradient aspects of these properties. Instead of focusing on the deviation of child language from the adult form, the variational model interprets child errors as the attestation of non-target but UG-possible grammars or hypotheses (Pinker, 1984, Crain, 1991). Specifically, the learner associates probabilities with the grammatical hypotheses made available by UG. When a learner hears an input token, it probabilistically selects a grammar or a parameter value to analyze this token. If successful, the selected grammar is rewarded and its probability goes up; if the grammar fails, its probability is penalized or lowered. This type of learning scheme is very general, falling in the tradition of Reinforcement learning in psychology and computer science, and has been identified across domains and species.

In Legate and Yang 2007, we develop a variational learning account of RI. We attribute children's nonfinite verbs to the type of UG-available grammar that does not mark tense on its root verbs, as attested for example in Mandarin. We call this the [-T] grammar, as contrasted with the [+T] grammar which does require tense marking on root verbs. If the English-learning child hears *I want water*, she cannot be sure that she is not learning a [-T] language, since the verb form of 1st person singular present is indistinguishable from the bare nonfinite form. On the other hand, if the English-learning child hears *I wanted water*, with overt morphological expression of past tense, she has reason to believe she is learning a [+T] language. Likewise, certain agreement verb forms (e.g., *John drinks coffee*), while not marking tense explicitly, is dependent on the presence of tense (secondary exponence; Carstairs 1987, Harley and Noyer 1999) thereby also providing evidence for the [+T] grammar. Given the abundance of you and me, and here and now in child directed speech, however, a significant proportion of root verbs that English-learning children hear do not show any evidence for tense marking. Hence, the child must rely on the morphological evidence for tense marking to learn that her language is [+T] and thereby unlearn the RI pattern for her language. Under the variational model, the speed with which the target grammar rises to dominance is determined by the frequency of disambiguating evidence that penalizes its competitor: given that languages offer different amounts of morphological evidence for tense marking, we predict that learners have different durations of the RI stage. Indeed, we found a significant inverse correlation between the amount of evidence for the [+T] grammar from tense-marked verbal inflections in the child-directed input data, and the duration of the RI stage across languages. Subsequent research has independently replicated our numerical results and, in some cases, found individual level correlation between the amount of tensed input and the rate of RI use between specific mother-child dyads (e.g. Rispoli et al. 2009).

It seems beyond doubt that the statistical properties of the input play an important role in the acquisition of language. It must be acknowledged that child language research in the generative tradition has rarely examined the quantitative aspects of linguistic input, which has given rise to the impression that input effects (such as the role of frequency) are inherently incompatible with the notion of UG and generative grammar. According to the alternative usage based approach, the child learner stores and retrieves lexically specific expressions in adult speech (Tomasello, 2000, 2003). Consider an alternative approach dubbed "MOSAIC" (e.g. Croker et al. 2000), which treats RI as the effect of purely input driven processing and learning. MOSAIC holds that the child stores lexically-specific combinations of words and other linguistic items, without the use of an abstract grammar. The child processes speech from right-to-left, retaining ever longer segments of the sentence: a sentence such as *She will jump* is retained as *Jump*, resulting in a root infinitive. Recent instantiations of this model (e.g. Freudenthal et al. 2010) have added an additional weaker left-to-right processing, in face of the obvious problem that early child speech does include initial wh-phrases and subjects, which would not be found through right-to-left processing. On a MOSAIC approach, there is no overarching abstract grammar and child language directly manifests the statistical composition of the input data. On the variational learning approach, the effects of the input are integrated into child language through abstract properties of the grammar. These differences lead to a stark contrast in predictions. For a variational learning model, verb forms that do not mark past tense, but rather mark agreement that is dependent on tense, do provide evidence for a [+T] grammar (as detailed above) and thus contribute to the child's successful acquisition of [+T], for which past tense is a specific realization. In other words, hearing *She is jumping* or *She jumps* assists the child in her acquisition of *She jumped*. For a MOSAIC learning

model, in contrast, agreement that is dependent on tense cannot be relevant to the learning of past tense – hearing *She is jumping* would allow the child to store *(She) jumping* and *She jumps* would allow the child to store *(She) jumps*, but neither results in storage of *(She) jumped*.

Given this background, we turn in the following section to our test case: the acquisition of past tense in African American English (AAE) and Mainstream American English (MAE).

2 The Acquisition of Tense Across English Dialects

2.1 Tense in AAE

As is well known, AAE employs considerably less tense and agreement marking than MAE; see Green 2002 for review. The third person singular present *-s* is very frequently omitted, resulting in *She jump*. In addition, Labov's (1969) classic study established the structural properties of copula deletion in AAE, as in *She nice*, including the important demonstration that copula omission in AAE obeys the same structural conditions as the contraction process in MAE *She's nice*. Furthermore, AAE uses the bare form of *be* with a habitual aspect interpretation. All these verbal forms constitute evidence for a [-T] grammar and against the target [+T] grammar: AAE is superficially more like Mandarin than MAE. But these non-tense-marked forms are used in variation with tense marked forms; thus collectively AAE children sometimes receive evidence for a [+T] grammar, and sometimes receive evidence for a [-T] grammar. Compared with MAE, however, these properties of AAE reduce the amount of [+T] evidence available to the AAE learning child and increase the amount of [-T] evidence.

For simple past tense, however, the two dialects do not differ: AAE does mark past tense: *She jumped*. This fact is verified quantitively in our analysis of child-directed AAE data presented below. Since AAE and MAE both mark past tense consistently, the grammar-based variational learning model and the usage-based MOSAIC make different predictions. On a MOSAIC model, the child carries out lexical learning by tracking and storing the usage patterns in adult speech. Therefore, we do not expect to find significant differences across dialect groups in child past tense marking, since the mothers use a comparable proportion of past tense tokens among their verbal forms. On a variational learning model, on the other hand, when the child learns tense, she is learning an overarching property of the language, [+/-T]. The MAE-learning children receive overall more evidence for [+T] marking than the AAE-learning: even though the past tense amounts are comparable, the MAE learner receives a good deal more evidence from third person singular, auxilaries and copulas, which are frequently absent in AAE. We therefore predict that MAE children use past tense more consistently than AAE children, despite hearing a comparable amount of past tense data in the input.

2.2 The Acquisition Data

Our study is based on the Hall corpus (Hall et al. 1984) in the CHILDES database (MacWhinney 2000). We used the data from 35 children between the age of 4;6 to 5;0 and their mothers (four children from the Hall corpus provided very few data points in the recording sessions and were excluded). The Hall corpus consists of four demographic groups: black working (BW) class, black professional (BP) class, white working (WW) class and white professional (WP) class.

To determine children's usage of past tense in obligatory contexts, we examined children's utterances by hand. We judged the conversational context to determine whether the context requires the use of past tense, and then recorded whether the child had indeed used the past tense. This gives us the rate of past tense usage for every verb for the children. For the mothers' data, we used the part of speech tagging and associated pattern-extraction tools used in the Legate and Yang 2007 study. The estimation of the amount of data in favor of the [+T] grammar from that study is consistent with results from an independent manual analysis of English data (Rispoli et al., 2009).

Table 1 gives a summary of the main results:

group	child past %	past tense % in CDS	all tense % in CDS
BW	87.2***	19.9	40.6***
BP	93.7	22.6	55.4
WW	94.4	17.1	49.8
WP	97.4	19.9	54.7

Table 1: Tense and past tense in child and mother's language. The first column provides children's percentage of past tense production in obligatory contexts. The second provides mothers' percentage of past tense forms in all utterances that contain a matrix verb. Finally, the third column provides mothers' percentage of forms providing evidence for a [+T] grammar in all utterances that contain a matrix verb (including past tense, and tense-dependent agreement).

Three notable findings emerge. First, there is no statistically significant difference in the percentage of utterances containing past tense forms over all matrix verb forms in the mothers across all four groups (column 2; $p = 0.89$). This confirms the observation that both AAE and MAE mark past tense consistently. Second, the BW class children produced a significantly lower rate of past tense than the other three groups of children ($p < 0.001$), which show no significant difference between them. Third, the BW class mothers produce a significantly smaller percentage of evidence for overall tense marking than the other three groups ($p < 0.001$); this reflects the well known differences between these dialects that we reviewed earlier. It also suggests that for the black professional class mothers, the dialect features of AAE with respect to tense marking are not very strong, and are in fact indistinguishable from the mothers in the white families in the data collected in the Hall corpus.

2.3 Input Mismatches and the Role of Grammar

Clearly, the differences in past tense marking between AAE working class children and the other three groups of children cannot be due to their mother's use of past tense per se, since there is no difference between the mothers' past tense usages. A series of correlation studies make the point more saliently.[1]

Figure 1 plots the percentage of past tense forms in a mother's speech against her child's past tense usage rate in obligatory contexts.

[1]For brevity, we only report the average rates of past tense marking for the 36 children. In the longer study under preparation, we provide results from mixed effects models with individual verbs and children as random effects. Our main conclusion is statistically confirmed: the rate of past tense marking in the children's language is not determined by the rate of past tense marking in the mothers' language, but rather by the overall evidence for tense marking in the mothers' language, especially third person singular.

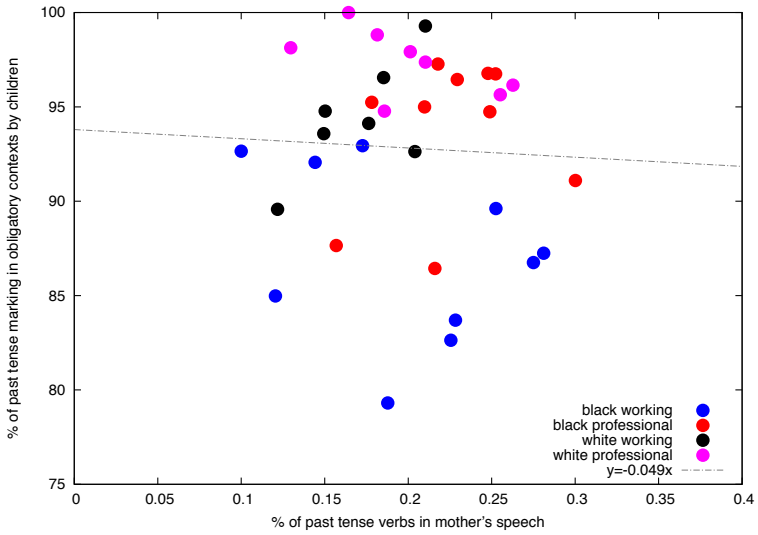

Figure 1: Past tense marking in child and child directed language.

Notably, the past tense usage shows *no* correlation at all between the input and output as can be seen in the linear regression line in Figure 1. We conclude from this that the child cannot be performing lexical learning in past tense; the MOSAIC model, which stores (sub)strings from the input, cannot predict a difference between AAE working class children and the other three groups, because the past tense input data for all the children are not statistically different.

The overall tense marking evidence, by contrast, strongly predicts with children's past tense usage ($r = 0.594, p < 0.001$). In fact, there is stronger correlation still between children's past tense usage and the percentage of *third person singular* forms in the input, with a correlation coefficient of 0.68. Figure 2 illustrates the result of the linear regression.

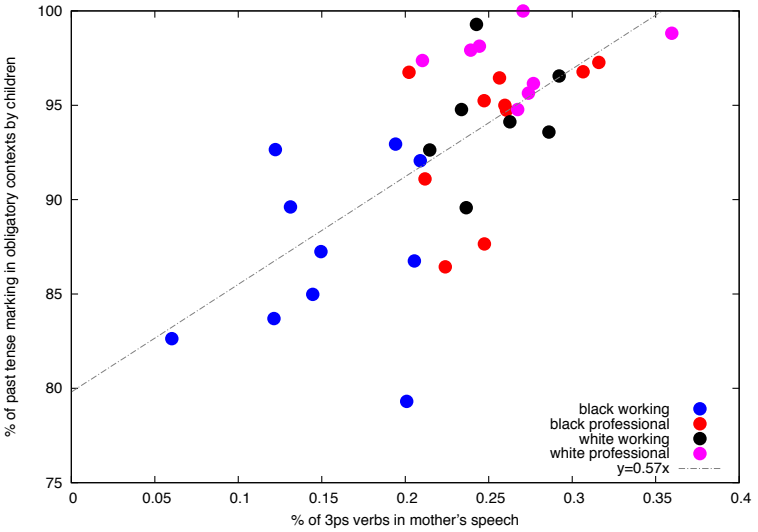

Figure 2: Past tense marking in child language and 3rd person singular marking in child directed language.

This demonstrates that the acquisition of past tense usage by children receives a boost from the overall amount of evidence for tense marking, even though the forms are not past tense per se. The effect of 3rd person singular in the input on children's past tense is expected on the variational learning approach: since there is no difference in past tense marking across the input groups, third person singular is among the chief contributors to the cross dialect differences in the rate of overall tense marking, and hence the stronger predictor.

In the spirit of the discussion in *Aspects*, the child learner assigns a structural description to the input and the grammatical feature of [+T] is activated and reinforced whenever any tensed form is encountered. However, this connection between third person singular in the input and past tense production is entirely mysterious under the usage-based learning account, where the input is viewed as a string of lexical items. Under the grammar-based account, by contrast, they are connected, and work collectively, toward the learning of tense in the language.

3 Summary

We conclude with some general remarks on the acquisition of the tense and the role of the linguistic input in language acquisition.

First, in Legate and Yang 2007, we acknowledged that while analysts can count the input frequency of various forms, we had little understanding how the child learns these morphological forms, their constitutive parts, as well as their corresponding syntactic and semantic features and properties. We are in slightly better position now (see Lignos and Yang, 2010), but much remains unclear.

Second, our study reinforces the grammar-based approach to tense and RI, showing that tense is an overarching and systematic property of the grammar that is not acquired in a piecemeal fashion. The variational learning model, which makes use of very general probabilistic learning schemes, offers a general framework for evaluating the effect of input frequencies interacting with the grammatical model under assumption. When a correct model of the grammar is assumed, one which assigns the proper structural description to the linguistic data, the statistical correlations–as well as non-correlations–between child language and adult language follow. When a wrong model of grammar is assumed, or no model of grammar at all, they do not.

Finally, we hope that our project brings generative researchers' attention to more quantitative studies of language variation and acquisition. Universal Grammar is compatible with input and frequency effects, which provide some of the best evidence for its validity. In addition to the inherent value in the study of dialects, important theoretical questions can be fruitfully studied and perhaps even resolved (Labov, 1969:761).

References

Carstairs, Andrew. 1987. *Allomorphy in inflexion*. London: Croom Helm.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge: MIT Press.

Chomsky, Noam. 1975. *Reflections on language*. New York: Pantheon.

Crain, Stephen. 1991. Language acquisition in the absence of experience. *Behavioral and Brain Sciences* 14:597–612.

Croker, Steve, Julian Pine, and Fernand Gobet. 2000. Modelling optional infinitive phenomena: A computational account of tense optionality in children's speech. In *Proceedings of the 3*rd

International Conference on Cognitive Modelling, 78–85.
Freudenthal, Daniel, Julian Pine, and Fernand Gobet. 2010. Explaining quantitative variation in the rate of optional infinitive errors across languages: A comparison of mosaic and the variational learning model. *Journal of Child Language* 37:643–669.
Green, Lisa J. 2002. *African American English: a linguistic introduction*. Cambridge University Press.
Hall, William S., William E. Nagy, and Robert Linn. 1984. *Spoken words: Effects of situation and social group on oral word usage and frequency*. Hillsdale, NJ: Lawrence Erlbaum.
Harley, Heidi, and Rolf Noyer. 1999. Distributed Morphology. *Glot International* 4:3–9.
Hyams, Nina. 1986. *Language acquisition and the theory of parameters*. Dordrecht: Reidel.
Jaeggli, Osvaldo, and Kenneth J. Safir. 1989. The null subject parameter and parametric theory. In *The null subject parameter*, ed. Osvaldo Jaeggli and Kenneth J. Safir, 1–44. Springer.
Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45:715–762.
Legate, Julie A., and Charles Yang. 2002. Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review* 19:151–162.
Legate, Julie A., and Charles Yang. 2007. Morphosyntactic learning and the development of tense. *Language Acquisition* 14:315–344.
Lignos, Constantine, and Charles Yang. 2010. Recession segmentation: Simpler online word segmentation using limited resources. In *Proceedings of the 14th Conference on Computational Language Learning*, 88–97.
MacWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum, 3rd edition.
Phillips, Colin. 1995. Syntax at age two: Cross-linguistic differences. *MIT Working Papers in Linguistics* 26:325–382.
Pinker, Steven. 1984. *Language learnability and language development*. Cambridge: Harvard University Press.
Rispoli, Matthew, Pamela A. Hadley, and Janet K. Holt. 2009. The growth of tense productivity. *Journal of Speech, Language, and Hearing Research* 52:930–944.
Roberts, Julie, and William Labov. 1995. Learning to talk Philadelphian: acquisition of short *a* by preschool children. *Language Variation and Change* 7:101–112.
Smith, Jennifer, Mercedes Durham, and Liane Fortune. 2009. Universal and dialect-specific pathways of acquisition: Caregivers, children, and t/d deletion. *Language Variation and Change* 21:69–95.
Tomasello, Michael. 2000. First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics* 11:61–82.
Tomasello, Michael. 2003. *Constructing a language*. Cambridge: Harvard University Press.
Valian, Virginia. 1991. Syntactic subjects in the early speech of American and Italian children. *Cognition* 40:21–81.
Yang, Charles. 2002. *Knowledge and learning in natural language*. Oxford: Oxford University Press.
Yang, Charles. 2004. Universal grammar, statistics or both? *Trends in Cognitive Sciences* 8:451–456.