On Perceived Conceptual and Methodological Divergences in Linguistic Theory and Cognitive Science: Distributional Analyses, Universal Grammar, and Language Acquisition

Garrett Neske

Many of the theoretical innovations in linguistic theory and cognitive science due to Noam Chomsky, and now considered to be cornerstones of the biolinguistic/*I*-language approach, are often viewed as radical replacements of earlier views in structural linguistics and psychology. In particular, the story often goes that Chomsky abandoned the structuralist tradition, along with "discovery procedures" and distributional analyses as a means toward understanding linguistic structure, in his development of transformational generative grammar. The position is also assigned to Chomsky that the study of language should consist of the elucidation of the innate component of the human mind and brain that allows for the acquisition of the rule systems of natural language and the real-time parsing of linguistic data. Both of these supposed truisms, however, require a serious reevaluation if the foundations of biolinguistics are to guide future research in the proper direction. Any divergences of generative grammar in linguistic theory and of universal grammar (UG) in cognitive science from previous conceptions of language and the mind have been too exaggerated. A careful consideration of the history and subsequent development of generative grammar and the biolinguistic/*I*-language approach will show, first, that distributional analyses were never abandoned in Chomsky's program and, second, that external linguistic data are integral to a theory of UG. These clarifications are absolutely essential if one is to make progress in biolinguistics.

Transformational generative grammar, as outlined in Chomsky (1957), detailed in Chomsky (1955, 1975a), and modified and refined several times throughout the second half of the last century (Chomsky 1965, 1972, 1975b, 1981, 1995) constituted a major departure from the structuralist tradition in linguistics. The aim of linguistic analysis from the structuralist, or distributionalist perspective, as exemplified in the work of Leonard Bloomfield (1933), Zellig Harris (1951), and others is to extract structure and constraints on linguistic representation from the distribution of formal objects such as phonemes. That is, the focus

I would like to thank Alec Marantz for his comments on an initial version of this paper and two anonymous reviewers for their indispensable comments and critique.

ISSN 1450–3417

should be on the likelihood of occurrence of a particular object given the occurrence of another object, with the aim of constructing hypotheses about the higher-level structure of language (Huck & Goldsmith 1986).

With the risk of oversimplifying, one may conceive of the distributional program as a bottom-up approach, akin to exploratory data analysis and clustering, in which the inference of structure comes from the distributional relations of primitive elements in a corpus, while Chomsky's transformational generative grammar, in contrast, takes a top-down approach, in which hypotheses about structure come from systems that generate a corpus. Working through many sentential examples, Chomsky became unconvinced that a distributional program could accomplish the task of inferring higher-level structure from the relations between elements. Chomsky's solution was a transformational generative grammar that could in principle generate the discrete infinity of grammatical expressions in a natural language. Furthermore, the 'transformational' component of this generative grammar differed markedly from the distributional program in positing different levels of representation connected by transformational rules. One level, deep structure in Chomsky's earliest Standard Theory (Chomsky 1957) and logical form in his most recent Minimalist Program (Chomsky 1995), maps syntactic structure to semantic interpretation and transformation rules use this level as input to yield an output of another level of representation, surface structure (Chomsky 1957) or phonetic form (Chomsky 1995).

In considering the birth of generative grammar and its subsequent development, one must be careful to consider the elements of earlier linguistic analysis that have been retained. Chomsky's generative approach was never completely divorced from distributionalism. Chomsky realized that purely distributional procedures could not, as an empirical matter, yield higher-level language structures from a corpus, which was the initial impetus for his construction of a top-down, meta-theory of language (Chomsky 1951). He did not, however, reject the distributional program altogether as a necessary component of the theory of grammar. In the years prior to the publication of Chomsky (1957), the new field of information theory introduced a principled, mathematical description of coding and transmission over a noisy channel (Shannon 1948), which, for Chomsky (1955, 1975a) and later for Harris (1991), seemed to bear on linguistic analysis. Shannon (1951) showed that information theory provided a way to predict letters in a text given the previous letters. Considering information-theoretic analysis as an empirical approach to the study of language structure, Chomsky notably worked with information theory pioneer Peter Elias to develop clustering algorithms for syntactic category formation and included the results of this analysis and a discussion of the scope and limits of distributional procedures in Chomsky (1955, 1975a: chap. V in particular). While Chomsky (1955, 1975a) is primarily an argument for a generative theory of language, there are several instances in which it is clear that a wholesale rejection of distributional, bottom-up methods in linguistic analysis is not the correct position to take, and in fact that a distributional approach might be indispensable for syntactic category formation:

> Note that there is no question being raised here as to the legitimacy of the probabilistic approach, just as the legitimacy of the study of meaning was in no way brought into question when we pointed out […] that projection can-

> not be defined in semantic terms. Whether or not the statistical study of language can contribute to grammar, it surely can be justified on quite independent grounds. These three approaches to language (grammatical, semantic, statistical) are independently important. In particular, none of them requires for its justification that it lead to solutions for problems which arise from pursuing one of the other approaches. (Chomsky 1975a: 148, fn. 19)

Thus, one would be too swift to regard Chomsky's approach as a successor to an extinct distributional program. Whether by lack of careful consideration of the foundations of generative theory or by trying to fit generative theory into some kind of Kuhnian paradigm shift, it is too often considered a truism that discovery procedures implemented to build categories and structures from primitive elements in linguistic data differ irrevocably from Chomsky's linguistics (Searle 1972). For one, the distributionalist program has been used to great effect by computational and corpus linguists for parsing, machine translation, and other tasks 'outside' of the biolinguistic/*I*-language approach. Harris's early hypothesis that morpheme boundaries corresponded to measurable variations in the complexity of phoneme sequences (Harris 1955) can be tested on a corpus of utterances (e.g., Hayes & Clark 1970, Tanaka-Ishii & Jin 2006), with the prediction that local phonemic entropy maxima will occur at morpheme boundaries. Yet, the most important point is that the distributional approach is not irrelevant to *I*-language, in which it is often assumed that the only relevant theory for language structure is generative and syntacto-centric (Jackendoff 1998). While distributionalist methods might fail at inducing syntactic structure, they are useful, and no doubt essential, to the segmentation problem (i.e. the determination of syntactic categories from linguistic data).

As a case in point, the famous example sentence in Chomsky (1955, 1957, 1975a), *Colorless green ideas sleep furiously*, is often touted as emblematic of the failure of distributional methods to induce structure from linguistic data. To a certain extent, this is true; the probability that each word in this sentence follows the other is thought to be vanishingly small, such that a parser would classify such a sentence as ungrammatical. Nevertheless, while the probability of the sentence occurring as written might be the same as that of the 'word salad' one obtains in reading it backwards, the sentence is pronounced with normal intonation and judged as grammatical by virtue of being in the class of sentences of the form Adjective–Adjective–Noun–Verb–Adverb (Chomsky 1975a: 145–147). A generative theory of language would require the segmentation of linguistic data into higher-level categories that serve as elements of the syntactic structure. That certain, modern distributional methods lead to a high probability of occurrence of *Colorless green ideas sleep furiously* compared to the reverse string (e.g., Pereira 2000) is actually consistent with the research program discussed in Chomsky (1955, 1975a); while distributional methods are not sufficient to induce syntactic structure, they are indispensable for the delineation of syntactic categories in linguistic data.

The birth of transformational generative grammar was intimately tied to the so-called cognitive revolution that rejected behaviorism (Chomsky 1959) and emphasized the role of innate cognitive architecture in producing behaviors. The position of *nativism* is often associated with the logical problem of language

acquisition and the argument from the poverty of the stimulus: How does the child learn the grammar of its native language from degenerate and limited linguistic data without a pre-existing structure specialized for the task? The guiding principle for understanding this structure is the notion of UG, a theory of the biologically instantiated ability to acquire and utilize the rule systems of a natural language (Chomsky 1965, 2006).

Given that linguistic structures exhibit a complexity not observed in any non-human communication system, or in any other behavior for that matter, it is almost obligatory to infer that there is an aspect of human biology and cognition, not shared with any non-human, responsible for the potential for acquisition of these structures and their use in cognition and communication. UG is a system that describes this unique ability. What kind of 'system' is UG posited to be? In the biolinguistic/*I*-language approach, UG is not a set of linguistic principles external to the individual like the grammars of specific natural languages, but an internal system for acquiring any natural language. In another sense, UG can be considered the initial state of the language faculty before any language-specific input. Thus, the focus of the biolinguistic/*I*-language approach is on 'human language' as opposed to the study of specific natural languages and associated corpora (i.e. *E*-language). This point is central to the later Principles–and–Parameters (P&P) approach to language acquisition (Chomsky & Lasnik 1993). The P&P approach holds that the individual has instantiated knowledge (though of course not 'conscious' knowledge) of fundamental principles of linguistic operation that are necessary components of all natural languages and that natural languages manifest their differences by setting parameters in the existing UG (i.e. the initial state of the language faculty).

Unfortunately, as in the perceived distributional/generative divergence in the analysis of linguistic structures, a similar artificial rift is drawn between UG and statistical learning, the build-up of language structure from the statistical properties of a sound signal (Seidenberg *et al.* 2002). In fact, these perceived divergences are both undoubtedly derived from the urge to fit differing conceptions of language analysis and language acquisition into an internalist–externalist, or even nature–nurture, debate. As before, a careful consideration of the specific claims immanent in the biolinguistic/*I*-language approach will reveal the exaggeration that statistical learning is somehow an alternative account of language acquisition that is in conflict with UG.

To reiterate, UG is a theory of the initial state of the language faculty, which, in the P&P model, undergoes a setting of parameters driven by external linguistic data. This is a selectionist account of learning that, while a relatively recent viewpoint in cognitive science, has been prevalent in the study of the development of biological structures (Piattelli-Palmarini 1989). There is noticeable confusion in the literature, and thus often artificial criticisms, about the claims of UG and the biolinguistic/*I*-language approach. For one, UG does not entail the 'unlearnability' of language, at least not in the selectionist sense, as has unfortunately been the position assigned to it in the literature (e.g., Bates & Elman 1996, Seidenberg 1997). UG is surely based upon the argument from the poverty of the stimulus, appealing to an innate structure to compensate for the lack of evidence needed to acquire the rule systems of natural language, but this does not relegate

the stimulus to negligible status. While children are theorized to be born with a UG, linguistic stimuli must allow the UG to converge on the correct grammar.

Many arguments countering, or purporting to counter, Chomsky's notion of UG and the biolinguistic/*I*-language approach rely on the efficacy of statistical learning in segmentation of the sound signal in early cognitive development. The landmark study, most often cited in statistical theories of language acquisition, demonstrating the power of statistical learning in the segmentation problem is that of Saffran *et al.* (1996). In this study, the authors show that infants can learn the 'words' of an artificial grammar within only minutes of exposure to sound samples of this grammar. They hypothesize that word segmentation is computed from local minima in transitional probabilities (TP) between syllables, where TP(A→B) = P(AB)/P(A) with P(AB) the probability of syllable B following A and P(A) the total frequency (i.e. probability) of A in the corpus (Yang 2004). While the data from Saffran *et al.* (1996) and others establish the dexterity with which children and adults detect the statistical distribution of sounds, it is far from straightforward what this implies about the innate capacity to learn language, let alone that a system like UG may not be necessary, which some authors have suggested (Seidenberg 1997, Bates & Elman 1996). In fact, it appears more likely that statistical learning does not play a role in deriving structure, but rather provides the language learner with a way to establish appropriate segmentations of the auditory signal, and thus the correct representation of linguistic elements, yet without the assignment of structure (Peña *et al.* 2002, Endress *et al.* 2005). The relation of these arguments to those asserting the divergence between the generative and distributional approaches to language analysis is evident (Gleitman 2002, in fact, argues that distributional approaches, while having little to do with cognitive science traditionally, have nonetheless been coapted to the field of language acquisition), and in both cases the same mistake is made; in the generative/distributional case, it is maintained that transformational generative grammar has no use for bottom-up procedures working on external data, while in the UG/statistical learning case, it is maintained that UG does not require any demarcation of the relational properties in the sound signal.

Statistical learning would, at first glance, appear to refute the need for a system like UG and validate a general-purpose learning scheme. The fact is, however, that there are an infinite number of statistical items in a sound signal of which the learner can keep track. The learner must be able to attend to the significant statistical correlations, such as transitional probabilities, and not, for instance, the probability of one syllable rhyming with the next or the probability that two adjacent vowels are both nasal (Yang 2004). This requires that the learner be equipped with some structure prior to linguistic exposure that accounts for the bias toward certain aspects of auditory stimuli and the neglect of other aspects. In effect, most arguments against the existence of a system like UG implicitly assume some initial structure that facilitates or constrains statistical learning. The appropriate question is not whether UG or a system like it exists, but what are its properties such that the learner can attend to certain features of the sound signal and acquire a complex rule system. General statistical learning schemes are not alone sufficient to account for the acquisition of a grammar. There is a need for an internal structure to attend to certain features of auditory stimuli.

Besides engaging in the segmentation problem, it might be that statistical learning can play a role in convergence to the correct grammar during language acquisition, though this has not yet been investigated experimentally like the segmentation problem (see Kuhl & Rivera-Gaxiola 2008 for a recent review of feasible experimental techniques in language acquisition). One issue in UG that must be addressed is that of parameter setting: What exactly does it mean for external linguistic data to 'set' a 'parameter'? The answer might involve a probabilistic component. In one learning scheme, UG represents the hypothesis space of grammars and parameter setting would involve the discarding of hypotheses that are inconsistent with external linguistic data. This notion is problematic, however, given the developmental data, which suggest that parameter setting is gradual and not punctuational (Bloom 1993) and that at any given time, the child has a representation of many possible grammars, not just one (Crain & Pietroski 2001). A scheme consistent with these data would be for each grammar in the hypothesis space to have a probability that is either increased or decreased depending on consistency or inconsistency with the linguistic input (Yang 2004). This kind of probabilistic hypothesis testing has been proposed as a neural basis for certain decision tasks (Gold & Shadlen 2002). While this situation still involves UG, statistical learning plays dual roles: segmentation of the sound signal and gradual convergence to the correct grammar. Again, to say that UG is incompatible with statistical learning is severely mistaken in the same way as the claim that transformational generative grammar is incompatible with the distributional approach and does not incorporate any of its methods.

It is important for the advancement of the biolinguistic/*I*-language approach not to overplay the supposed methodological and conceptual distance between distributionalism/statistical learning on the one hand and transformational generative grammar/UG on the other. All of these approaches are necessary for a full account of the structure of natural language and the development of grammar in the individual. Without distributional methods, for instance, an objective account of the notions of 'word' or 'lexicon', both essential to transformational generative grammar (e.g., the Merge operation acting upon *lexical* items in the Minimalist Program of Chomsky 1995), would be severely lacking. On the cognitive-science side, while UG is viewed as a biologically instantiated template to which the grammar of a natural language must conform, the theory of UG and the P&P approach have very little to say about the mechanisms of this 'conforming'. As in the necessity of the distributional approach for inducing syntactic categories, statistical learning appears sufficient (necessary?) to segment the sound signal into primitive linguistic elements based upon the statistics inherent in the stimulus. The propensity to frame new ideas as revolutionary attacks against past research programs is strong, but must often be mitigated through a careful consideration of the claims inherent in these new ideas and the extent to which they build upon previous ones. With this prescription in hand, it should be clear that the transformational generative grammar/distributionalism and UG/statistical learning chasms are not as wide as they are purported to be. In fact, all of these concepts and methods are utilized to some extent in the modern biolinguistic/*I*-language approach to natural language structure and the development and use of the biological language faculty.

References

Bates, Elizabeth & Jeffrey Elman. 1996. Learning rediscovered. *Science* 274, 1849–1850.

Bloom, Paul. 1993. Grammatical continuity in language development: The case of subjectless sentences. *Linguistic Inquiry* 24, 721–734.

Bloomfield, Leonard. 1933. *Language*. New York: Henry Holt.

Chomsky, Noam. 1951. Morphophonemics of Modern Hebrew. Philadelphia, PA: University of Pennsylvania MA thesis.

Chomsky, Noam. 1955. *The Logical Structure of Linguistic Theory*. Cambridge, MA: Microfilm, MIT Humanities Library.

Chomsky, Noam. 1957. *Syntactic Structures*. The Hague: Mouton.

Chomsky, Noam. 1959. A Review of B.F. Skinner's *Verbal Behavior*. *Language* 35, 26–58.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1972. *Studies on Semantics in Generative Grammar*. The Hague: Mouton.

Chomsky, Noam. 1975a. *The Logical Structure of Linguistic Theory*. New York: Plenum Press.

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

Chomsky, Noam. 1981. *Lectures on Government and Binding: The Pisa Lectures*. Dordrecht: Foris.

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, Noam. 2006. *Language and Mind*, 3rd edn. Cambridge: Cambridge University Press.

Chomsky, Noam & Howard Lasnik. 1993. The theory of principles and parameters. In Joachim Jacobs, Arnim von Stechow, Wolfgang Sternefeld & Theo Vennemann (eds.), *Syntax: An International Handbook of Contemporary Research*, vol. 1, 506–569. Berlin: Walter de Gruyter.

Crain, Stephen. & Paul Pietroski. 2001. Nature, nurture and universal grammar. *Linguistics and Philosophy* 24, 139–186.

Endress, Ansgar D., Brian J. Scholl, & Jacques Mehler. 2005. The role of salience in the extraction of algebraic rules. *Journal of Experimental Psychology: General* 134, 406–419.

Gleitman, Lila R. 2002. Verbs of a feather flock together II: the child's discovery of words and their meanings. In Bruce E. Nevin & Stephen M. Johnson (eds.), *The Legacy of Zellig Harris: Language and Information into the 21st Century*, vol. 1, 209–229. Philadelphia: John Benjamins.

Gold, Joshua I. & Michael N. Shadlen. 2002. Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron* 36, 299–308.

Harris, Zellig S. 1951. *Methods in Structural Linguistics*. Chicago: University of Chicago Press.

Harris, Zellig S. 1955. From phoneme to morpheme. *Language* 31, 190–222.

Harris, Zellig S. 1991. *A Theory of Language and Information: A Mathematical Approach*. Oxford: Oxford University Press.

Hayes, John R. & Herbert H. Clark. 1970. Experiments on the segmentation of an artificial speech analogue. In John R. Hayes (ed.), *Cognition and the Development of Language*, 221–234. New York: John Wiley and Sons.

Huck, Geoffrey J. & John A. Goldsmith. 1986. Distributionalist and mediationalist themes in the development of linguistic theory. *Linguistic Inquiry* 17, 265–299.

Jackendoff, Ray. 1998. The architecture of the language faculty: A neominimalist perspective. In Peter Culicover & Louise McNally (eds.), *The Limits of Syntax*, 19–46. New York: Academic Press.

Kuhl, Patricia & Maritza Rivera-Gaxiola. 2008. Neural substrates of language acquisition. *Annual Review of Neuroscience* 31, 511–534.

Pereira, Fernando. 2000. Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society A* 358, 1239–1253.

Peña, Marcela, Luca L. Bonatti, Marina Nespor & Jacques Mehler. 2002. Signal-driven computations in speech processing. *Science* 298, 604–607.

Piatelli-Palmarini, Massimo. 1989. Evolution, selection and cognition: From "learning" to parameter setting in biology and in the study of language. *Cognition* 31, 1–44.

Saffran, Jenny R., Richard N. Aslin & Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274, 1926–1928.

Searle, John R. 1972. Chomsky's revolution in linguistics. *The New York Review of Books* 18 (June 29), 16–24.

Seidenberg, Mark S. 1997. Language acquisition and use: learning and applying probabilistic constraints. *Science* 275, 1599–1603.

Seidenberg, Mark S., Maryellen C. MacDonald & Jenny R. Saffran. 2002. Does grammar start where statistics stop? *Science* 298, 553–554.

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27, 379–423, 623–656.

Shannon, Claude E. 1951. Prediction and entropy of printed English. *Bell System Technical Journal* 30, 50–64.

Tanaka-Ishii, Kumiko & Zhihui Jin. 2006. From phoneme to morpheme: Another verification using a corpus. *Lecture Notes in Computer Science*, 4285, 234–244.

Yang, Charles D. 2004. Universal grammar, statistics or both? *Trends in Cognitive Sciences* 8, 451–456.

Garrett Neske
New York University
Department of Linguistics
10 Washington Place
New York, NY 10003
USA
gtn206@nyu.edu