ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 2, No. 9, pp. 1917-1921, September 2012

doi:10.4304/tpls.2.9.1917-1921

Research on Statistical Mechanisms of Language Acquisition

Haiyan Han
School of Foreign Languages, University of Ji'nan, Ji'nan, Shandong, China
Email: sfl_hanhy@ujn.edu.cn

***Abstract*—What mechanisms underlie human language acquisition? Relevant evidence indicates that language learners, with infants included, have the ability to employ statistical properties of linguistic input to find language structure, such as sound patterns, vocabulary, and grammar. These abilities appear to be both powerful and constrained, so that a certain number of statistical patterns are more easily mastered and employed than others. Implications for the structure of human languages are to be discussed.**

***Index Terms*—statistical properties, language acquisition, mechanisms**

I. Introduction

Suppose one is confronted with the following challenge: He must find the underlying pattern of a system that includes thousands of pieces, all generated by combing a small group of elements in different ways. These pieces, in turn, can be combined in an unlimited number of ways, although only a subset of those combinations is actually acceptable. However, the subset that is acceptable is itself infinite. He somehow should quickly find out the pattern of this system so that he can employ it correctly early in his childhood.

This system mentioned above definitely is human language. These elements are the sounds of language, and the larger pieces are the words, which in turn combine to form sentences. Considering the variety and complexity of human language, it seems impossible that children could ever discover its structure. The process of acquiring such a system seems to be almost as complex as the system itself, so it is not surprising to notice that the mechanisms underlying language acquisition are a matter of much debate. This debate focuses on the innate and environmental contributions to the process of language acquisition, and the degree to which these components draw on information and abilities that are also related to other domains of learning.

Particularly, there exists a fundamental conflict between theories of language acquisition in which learning plays an important role and theories in which learning is considered unimportant. A strong point of learning-oriented theories is that they use the increasing evidence suggesting that young humans possess powerful learning mechanisms. For example, infants can quickly use the statistical properties of their language environments, including the distributions of sounds in words and the orders of word types in sentences, to discover important elements of structure. Infants can track such statistical properties, for example, to find speech categories (e.g., native language consonant; e.g. Maye, & Gerken, 2002), word boundaries (e.g., Saffran, Aslin, & Newport, 2001b), and rudimentary syntax (Gomez & Gerken, 1999).

However, language acquisition theories believing that learning plays a crucial part attract much negative comments. One of the most prevalent negative comments against learning-oriented theories is that such accounts seem contradictory to one of the central observations about human languages. Although there are surface differences between the linguistic systems of the world, they share a lot of similarities. They vary in non-arbitrary ways. Theories of language acquisition that mainly focus on inborn knowledge of language do give a sufficient explanation for cross-linguistic similarities. The seminal work of Noam Chomsky is one of such theories, suggesting that universal grammar is pre-set in the children's linguistic endowment, and do not require specific learning. Such explanations lead to predictions about the types of patterns that should be observed cross-linguistically, and result in important claims regarding the evolution of a language ability that includes innate knowledge (Pinker & Bloom, 1990).

Can the learning-based theories also explain the existence of linguistic universal grammar? The answer to this question is the focus of current research. The constrained statistical learning framework indicates that learning is crucial to language acquisition, and that the fundamental nature of language learning accounts for the cross-linguistic similarities. The key point is that learning is constrained, and learners are not open-minded. Language learners calculate some statistics more easily than others. What interests researcher most is those constraints on learning that are related with similarities between human languages (Newport & Aslin, 2000). According to this framework, cross-linguistic similarities are indeed not accidental, as is suggested by the framework of Chomsky. But they are not the result of innate linguistic knowledge. Instead, human languages have been to some degree determined by human learning mechanisms (together with constraints on human perception, processing, and speech generation), and aspects of language that improve learnability are apt to persist in linguistic structure than those that do not. Consequently, according to that point of view, cross-linguistic similarities are not the result of innate knowledge, as is traditionally believed, but rather are the result of constraints on learning. Moreover, if human languages were not determined by

constraints on human learning mechanisms, it is likely that these learning mechanisms and their constraints were not modified only for language acquisition. Instead, learning in non-linguistic domains should be similarly constrained, as in fact seems to be the case.

A deeper understanding of these constraints may result in new connections between theories that concentrate on nature and theories that concentrate on nurture. Constrain-oriented learning mechanisms require both specific experiences to promote learning and pre-existing structures to get and manipulate those experiences.

II. Grammar and Languages

A necessary preliminary distinction is between an 'I-language' (Internal) approach to grammar and an 'E-language' (External) approach, as Chomsky has termed it. An I-language approach concentrates on the knowledge of language stored in the mind of the individual—a system represented in the mind/brain of a particular individual; an I-language grammar tries to mirror this mental reality. An E-language approach on the other hand studies a collection of data separate from the speaker's mind; an E-language grammar describes the regularities and patterns found in the collection –a grammar is a collection of descriptive statements concerning the E-language. I-language grammars typically rely on example sentences; E-language grammars on transcripts of spoken language or written texts. The contrast is partly between a psychological approach that sees language as part of the individual mind and a sociological approach that sees it as part of the community. In a sense recent language teaching has concentrated on the E-language end –on 'behaviour' and 'ommunication'—rather than keeping a balance between I-language and E-language perspectives.

The grammar of a language is an account of the native speaker's knowledge of language. A speaker of English knows, for instance, that English declarative sentences usually have overt subjects and verb subject order; a native speaker of Spanish knows that such sentences need not have subjects and may have verb subject order as well as VS order. The language student is attempting to acquire some aspects of this knowledge. Hence the grammar plays some part in the description of what the student has to know, the syllabus.

To I-language theorists the grammar is also an account of what the native speaker has learnt. The language knowledge that is stored must have a source; syntax can be considered a description of what a human mind comes to know, given exposure to a human language. In the Universal Grammar theory, the description of language knowledge is in part an account of the principles of grammar that are already present in the mind waiting to be triggered; appropriate data pushes the child towards English, Spanish, or Chinese. Grammar is therefore needed as one strand in the student's acquisition of a new language.

The grammar is in addition a partial account of how the native speaker processes language. While grammar represents language in a static form, this representation is also related to the processes native speakers use in language comprehension and production. For example the Marcus parser (Marcus, 1980) and Augmented Transition Network parsers (Wanner and Maratsos, 1978) show how particular models of syntax can be used as models for language processing provided they are supplemented with plausible memory constraints on 'lookahead' or working memory. Inasmuch as language students are processing and learning to process language, such aspects of grammar are important both as their ultimate target and for immediate use in the classroom. Overall, grammar is important for language teaching as an account of part of the knowledge the students want to attain, and hence of what they have to learn, and as a partial account of the processes involved in language production and comprehension. This affects firstly the syllabus the teacher wants to use, which relates to the native speaker's knowledge; secondly the sequence for introducing elements the teacher adopts, which relates to the learning process; and thirdly the classroom techniques the teacher employs, which make use of language processes. Even if the overall goal of language teaching is confined to communication, grammar necessarily plays some part in each of these levels; applied linguists need to consider the relationship of current grammatical theories to each of them.

The definition of a grammar is central to most work in statistical linguistics and natural language processing. A grammar is a description of a language; generally it identifies the sentences in the language and describes them, e.g., by defining the phrases of a sentence, their relationships, and perhaps some aspects of their deep meanings. The formal framework, whether used in a generative grammar, or statistical linguistics is due to Chomsky.

If T is a finite set of symbols, let $T*$ be the set of all strings (i.e., finite sequences) of symbols of T, including the empty string, and let T+ be the set of all nonempty strings of symbols of T. A language is a subset of $T*$. A rewrite grammar G is a quadruple G = (T, N, S, R), where T and N are disjoint finite sets of symbols (called the terminal and non-terminal symbols respectively), $S \in N$ is a distinguished non-terminal called the start symbol, and R is a finite set of productions. A production is a pair (α, β) where $\alpha \in N+$ and $\beta \in (N \cup T)*$; productions are usually written $\alpha \rightarrow \beta$. Productions of the form $\alpha \rightarrow \varepsilon$, where ε is the empty string, are called epsilon productions.

A rewrite grammar G defines a rewriting relation $\Rightarrow_G \subseteq (N \cup T)* \times (N \cup T)*$ over pairs of strings consisting of terminals and non-terminals as follows: $\gamma\alpha\delta \Rightarrow \gamma\beta\delta$ iff $\alpha \rightarrow \beta \in R$ and $\gamma, \delta \in (N \cup T)*$ (the subscript G is dropped when clear from the context). The reflexive, transitive closure of $\Rightarrow$ is denoted $\Rightarrow*$. Thus $\Rightarrow*$ is the rewriting relation using arbitrary finite sequences of productions. (It is called "reflexive" because the identity rewrite, $\alpha \Rightarrow* \alpha$, is included). The language generated by G, denoted LG, is the set of all strings $w \in T+$ such that $S \Rightarrow* w$.

A terminal or non-terminal $X \in N \cup T$ is useless unless there are $\gamma, \delta \in (N \cup T)*$ and $w \in T*$ such that $S \Rightarrow* \gamma X \delta \Rightarrow* w$. A production $\alpha \rightarrow \beta \in R$ is useless unless there are $\gamma, \delta \in (N \cup T)*$ and $w \in T*$ such that $S \Rightarrow* \gamma\alpha\delta \Rightarrow \gamma\beta\delta \Rightarrow* w$.

Informally, useless symbols or productions never appear in any sequence of productions rewriting the start symbol S to any sequence of terminal symbols, and the language generated by a grammar is not affected if useless symbols and productions are deleted from the grammar (Chomsky, 1957).

III. Statistically-based Learning the Sounds of Words

In order to explore the nature of infants' learning mechanisms, my coworkers and I studied an aspect of language that we knew must surely be learned, namely, word segmentation, or the boundaries between words in fluent speech. This is much difficult for infants acquiring their first language, because speakers do not mark word boundaries with pauses, as is shown in Figure 1. Instead, infants must determine in where place one word ends and the next begins without access to obvious acoustic cues. Learning is required in this process because infants cannot innately know that, for example, *pretty* and *baby* are words, but *tyba* (spanning the boundary between *pretty* and *baby*) is not.

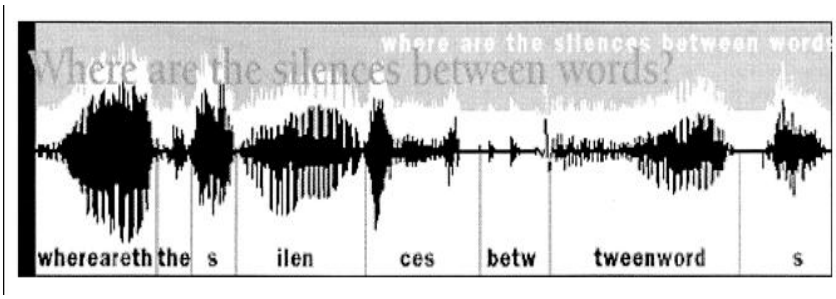

Figure 1. A speech waveform of the sentence "Where are the silences between words?"

In Figure 1, the height of the bars shows loudness, and the x-axis means time. This example demonstrates the shortage of consistent silences between word boundaries in fluent speech. The vertical gray lines stand for quiet points in the speech stream, some of which do not correspond to word boundaries. Some sounds are shown twice in the transcription below the waveform because of their continued persistence over time.

One source of information that may explain the existence of word boundaries is the statistical structure of the language in the infant's environment. In English, the syllable *pre* occurs before a small set of syllables, consisting of *ty*, *tend*, and *cedes*; in the speech stream, the probability that *pre* is followed by *ty* is thus greatly high (about 80% in speech to young infants). However, because the syllable *ty* occurs word finally, it can be followed by any syllable that can start an English word. Consequently, the probability that *ty* is followed by *ba*, as in *pretty baby*, is rather low (about 0.03% in speech to young infants). The difference in sequential probabilities proves that *pretty* is a word, and *tyba* is not. More generally, considering the statistical properties of the input language, the capacity to track sequential probabilities would be a useful tool for young learners.

To investigate whether human beings can use statistical learning to find word boundaries, we played adults, first graders, and 9-month-olds recordings of nonsense languages in which the only cues to word boundaries were the statistical properties of the sequential syllables. Listeners briefly heard a continuous sequence of syllables consisting of multisyllabic words from one of the languages (e.g., *golabupabikutut ibubabupugolabubabupu.*.). Then, the listeners were tested to determine whether they could distinguish the words from the language from sequences spanning word boundaries. For example, we made a comparison between performance on words like *golabu* and *pabiku* with performance on sequences like *bupabi*, which spanned the word boundaries. To fulfill the task, listeners would have had to track the statistical properties of the input. The test results proved that human learners, including infants, can indeed use statistics as a tool to find word boundaries. Additionally, this capacity does not just belong to humans: Cotton-top tamarins, a monkey species, can also track statistics to find word boundaries.

These results immediately raised one question about the degree to which statistical learning is constrained to language-like stimuli. A large number of results indicate sequential statistical learning is quite general. For example, infants can track sequences of tones, finding "tone-word boundaries" using statistical cues (Saffran, 2003), and can acquire statistically-oriented visual structures (Kirkham, Slemmer, & Johnson, 2002); work in progress is extending these results to the field of events in human action sequences.

Considering that the ability to find units via their statistical coherence is not confined to language (or to human beings), one might ask whether the statistical learning results actually apply to language at all. To put it another way, do infants actually acquire the real-world language using statistical learning mechanisms? One way to approach this question is to ask what infants are actually learning in our segmentation task. Are infants learning statistics? Or are infants learning language via statistics? Our results show that when infants being exposed to English-speaking environment have segmented the sound strings, they treat these nonsensical patterns as English words. Statistical language learning in the laboratory thus seems to be integrated with other aspects of language acquisition. Relevant results suggest that 11-month-olds can first segment novel words and then find syntactic regularities relating the new words—all the same group of input. This would not be possible if the infants built mental pictures only of the sequential probabilities relating individual syllables, and no word-level pictures (Saffran & Wilson, 2003). These findings reveal a

constraint on statistical language learning: The mental representations produced in this process are not just groups of statistically-linked syllables, but new groups that are available to serve as the input to subsequent learning process.

Similarly, it is possible to examine constraints on language learning that probably influence the acquisition of the sound structure of human languages. The types of sound patterns that infants learn most easily may be more dominant in languages than are sound patterns that are not learnable by infants. We tested the hypothesis by asking whether infants discover some phonotactic regularities (constraints on where particular sounds can appear; e.g., /fs/ can appear at the end, but not the beginning, of syllables in English) easier to learn than others (Saffran & Thiessen, 2003). The results reveal that infants readily learn novel regularities that are similar to the types of patterns in the world's languages, but cannot learn regularities that are not consistent with natural language structure. For example, infants quickly learn new phonotactic regularities about generalizations across sounds that share a phonetic feature, while cannot learn regularities that disregard such features. Therefore, infants can more easily learn a group of patterns that group together /p/, /t/, and /k/, which are all voiceless, and that group together /b/, /d/, and /g/, which are voiced, than to learn a pattern that group together /d/, /p/, and /k/, but does not apply to /t/. Such studies may give an explanation for why languages show the types of sound patterning that the do; sound patterns that are hard for infants to learn may be unlikely to occur across the languages in the world.

IV. Statistical Learning and Syntax

Issues about learning versus innate knowledge are most dominant in the field of syntax. How could learning-based theories explain the acquisition of abstract structure (e.g., phrase boundaries) not obviously mirrored in the surface statistics of the input? Unlike explanations focused on innate linguistic knowledge, most learning-oriented theories do not give a clear account of the ubiquity of particular cross-linguistic structures. One way to approach these issues is to ask whether some nearly universal structural aspects of human languages may be the result of constraints on human learning (Morgan, Meier, & Newport, 1987). To test the hypothesis, we asked whether one such aspect of syntax, phrase structure (groups of types of words together into sub-groups, such as noun phrases and verb phrases), is the result of a constraint on learning: Do humans acquire sequential structures more readily when they are grouped into subunits such as phrases than when the are not? We discovered a statistical cue to phrase units, predictive dependencies (e.g., the presence of a preposition like *the* or *a* predicts a noun somewhere in the following part; the presence of a preposition predicts a noun phrase somewhere in the following part), and determined that learners can use this kind of cue to discover phrase boundaries (Saffran, 2001a).

To directly test of the theory that predictive dependencies improve learnability, we made a comparison between the acquisition of two nonsense languages, one with predictive dependencies as a cue to phrase structure, and one with no predictive dependencies (e.g., words like *the* could appear either with or without a noun, and a noun could appear either with or without words like *the*; neither type of word predicted the presence of the other). We found that listeners learned language better when they were exposed to languages with predictive dependencies than when they were exposed to languages with no predictive dependencies (Saffran, 2001b). Much to our interest, the same constraint on learning occurred in tasks using nonlinguistic materials (e.g., computer alert sounds and simultaneously presented shape arrays). These findings proved the claim that learning mechanisms not specifically designed for language learning may have determined to some degree the structure of human languages.

V. Future Research

Results to date indicate that human language learners have powerful statistical learning abilities. These mechanisms are constrained at multiple levels; there are constraints on what information serves as input, which computations are performed over that input, and the pattern of the representations that occur as output. To better understand the contribution of statistical learning to language acquisition, it is necessary to determine the degree to which statistical learning explain given the complexities of the acquisition process. For example, how does statistical learning interact with other aspects of language acquisition? One solution to this question is to explore how infants weight statistical cues related to other cues to word segmentation early in life. The results of such studies give us an insight into the ways in which statistical learning may help infant learners to determine the relevance of the many cues inherent in language input. Similarly, we are studying how statistics meet up with meaning in the world (e.g., are in statistics defined "words" easier to learn as labels for novel objects than sound sequences spanning word boundaries?), and how infant learners in bilingual environments deal with multiple groups of statistics. Research on the interaction between statistical learning and the rest of language learning may give a better explanation of how various non-statistical aspects of language are acquired. Additionally, a better explanation of the learning mechanisms used successfully by typical language learners may help researchers better understand the types of processes that go awry when infants do not acquire languages as easily as their peers.

It is also crucial to find which statistics are available to children and whether these statistics are actually related to natural language structure. Researchers are divided in opinion on the role that statistical learning should play in acquisition theories. For example, they disagree about when learning is best explained as statistically based as opposed to rule based (i.e., using mechanisms that operate over algebraic variables to find abstract knowledge), and about

whether learning can still be regarded as statistical when the learning input is abstract. Debates over the proper role for statistical learning in language acquisition theories cannot be resolved in advance of the data. For example, though one can differ between statistical versus rule-based learning mechanisms, and statistical versus rule-based knowledge, the data are not yet enough to determined whether statistical learning renders rule-based knowledge structures, and whether abstract language knowledge can be statistically-based probablistic. More empirical hypotheses will be required to resolve these relevant theoretical disagreements.

Moreover, more investigations into humans and other species may give more explanatory powers with respect to the relationship between statistical learning and human language. Present research is identifying species differences in statistical learning mechanisms (Newport, 2000). Considering that nonhumans and humans track different statistics, or track statistics over different units, learning mechanisms that do not seem to be human-specific may in fact generate human-specific results. Alternatively, what the human learning mechanisms and nonhuman learning mechanisms share may indicate that differences in statistical learning cannot explain cross-species differences in language-learning abilities.

VI. Conclusion

Clearly enough, human language is a system of much complexity. Meanwhile, the use of statistical cues may be useful for language learners to find some of structures in language input. To what degree can the kinds of statistical structures accessible to language learners help in revealing the complexities of this system? Although the answer to this question is not known, it is possible that a mixture of innate constraints on the types of structures acquired by language learners, and the use of output from a certain level of learning as input to another, may be useful to account for why something complex is readily acquired by the humans. The language learning mechanisms may have played a crucial part in molding human language patterns.

References

[1] Chomsky, N. (1957). Syntactic Structures. Mouton: The Hague.
[2] Fiser, J., & R. N. Aslin. (2004). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, 4*, 127-134.
[3] Gomez, R. L., & L. Gerken. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 1*, 97-108.
[4] Kirkham, N. Z., Slemmer, J. A., & S. P. Johnson. (2002). Visual statistical learning in infancy: Evidence of a domain general learning mechanism. *Cognition, 2*, 114-139.
[5] Maye, J., Werker, J. F., & L. Gerken. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 3*, 224-245.
[6] Morgan, J. L., Meier, R. P., & E. L. Newport. (1987). Structural packaging in the input to language learning: Contributions of international and morphological marking of phrases to the acquisition of language. *Cognitive Psychology, 2*, 207-232.
[7] Newport, E. L., & R. N. Aslin. (2000). Innately constrained learning: Blending old and new approaches to language acquisition. In S.C. Howell, S. A. Fish, & T. Keith-Lucas (Eds.), *Proceedings of the 24th Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press, 167-197.
[8] Pinker, S., & P. Bloom. (1990). Natural language and natural selection. *Behavioral and Brain Sciences, 4*, 78-98.
[9] Saffran, J. R. (2001a). The use of predictive dependencies in language learning. *Journal of Memory and Language, 2*, 137-145.
[10] Saffran, J. R. (2001b). Words in a sea of sounds: The output of statistical learning. *Cognition, 3*, 221-265.
[11] Saffran, J. R., & D. P. Wilson. (2003). From syllables to syntax: Multi-level statistical learning by 12-month-old infants. *Infancy, 1*, 98-120.

Haiyan Han was born in Qitaihe of Heilongjiang Province, China in 1976. She received her Master's Degree in Foreign Linguistics and Applied Linguistics from University of Jinan, China in 2011.

She is currently a lecturer in the School of Foreign Languages of University of Jinan, Jinan, China. Her research interests include language teaching methodologies and applied linguistics.