Why sentences are more complex than words

Jeffrey Heinz[1*] and William Idsardi[2]

[1]Department of Linguistics and Cognitive Science, University of Delaware,
42 E Delaware Avenue, Newark, DE 19716, USA
[2]Department of Linguistics and Program in Neuroscience and Cognitive Science,
University of Maryland,
1401 Marie Mount Hall, College Park, MD 20742, USA

[*]To whom correspondence should be addressed; E-mail: heinz@udel.edu.

Differences in the computational complexity of syntax and phonology imply multiple, modular learning procedures within the domain of natural language.

Debates between domain-specific (*1*) and domain-general (*2*) learning strategies often focus on the possible existence of a module specialized for speech and language (*3*). Language perception is organized at different levels: the organization of sounds into words (phonology), the organization of roots and affixes into words (morphology), and the organization of words into phrases and sentences (syntax). Each of these areas has its own internal organizing principles. Some well-known and less-well-known differences in the computational complexity of syntax and phonology imply that humans exploit these differences with specialized learning mechanisms for syntax and phonology within the domain of language itself.

A significant, overlooked result is that natural language sound patterns are measurably less complex than word patterns. What are word and sound patterns? Well-formed and ill-formed patterns of words into sentences are familiar to everyone. Every sentence in this article is well-

formed; if every period is shifted one word to the left we obtain a set of ill-formed sentences. Sound patterns (*4*) are also familiar: English speakers realize that *Gdansk* and *srem* are not native English words (*5*) (though they are possible words in other languages) because they violate the proper sequencing of English sounds. As a result, English speakers readily assent to some new coinages (*bling*) while avoiding others (*gding*).

How is complexity measured in language? Mathematically, a language is a set of strings; i.e. sequences of more basic units (*6,7*). Sentences are sequences of words; words are sequences of sounds. Theoretical computer science provides a mathematically rigorous way to characterize sets of sequences in both language (*8*) and other fields (*9*). The Subregular and Chomsky hierarchies (*10, 11*) arrange patterns into nested regions (see Figure). It is a surprising result that many different grammar formalisms, including probabilistic versions (*12*), converge to the same areas of these hierarchies. Since any pattern can be described by a set of strings, or by a probability distribution over strings, a distinct advantage of this framework is it allows for the comparison of patterns in different domains.

When sentences are compared with words, sentences can require mildly *context-sensitive* computations (*13, 14*) whereas words never require more than *regular* computations (*15, 16*). For example, English contains recursive sentences because sentences can be embedded into larger ones, e.g. *The mouse <u>that the cat chased</u> ran away* (*8*). The recursive nature of sentences is a defining characteristic of natural language (*17*), and provably makes them at least context-free (*8*). In contrast, words are measurably simpler. Phonology not only restricts adjacent sounds (recall *gding*), but can also restrict consonants and vowels over long distances (*18, 19*). For example, Samala, a native language of North America, does not allow words which contain both ‘s’ and ‘sh’ (*20*). There are words like *<u>sh</u>toyonowonowa<u>sh</u>* ‘it stood upright’ but none like *<u>sh</u>toyonowonowa<u>s</u>* (*21*). Moreover, no known sound pattern is recursive. In fact, all can be described, including the long-distance ones, with regular grammars (*16*), and are probably

even less complex. Thus syntax is at or above context-free complexity, and phonology is below. This dramatic computationally measurable difference between sentences and words demands explanation.

One possibility is that sound sequences within words are constrained by psychophysical properties of the human nervous, motor, and auditory systems in ways that word sequences within sentences are not. That is, the moment-to-moment production of sounds is constrained by psychophysical constraints on the moment-to-moment configurations of the vocal tract, a theory known as co-articulation (*22*, *23*). There are two problems with this view. First, languages have different patterns of co-articulation (*24*) and so phonology cannot be reduced completely to psychophysics. Second, this hypothesis fails to account for long-distance patterns like the one in Samala for the simple reason that the tongue does not retain the "sh" posture throughout *shtoyonowonowash*. If co-articulation is abstracted into relative adjacency (i.e. making the "sh" sounds abstractly adjacent) then this abstraction abandons the psychophysical explanation, while keeping sound patterns in words *regular*.

Another possibility is that humans employ different learning mechanisms for phonology and syntax. There is a wide convergence of results from philosophy (*25*), computer science (*26–28*), and psychology (*1*) that learning is only possible if learners are restricted in the hypotheses that they are allowed to consider. In fact, the successes in what Meltzoff et al. (*29*) term 'the new science of learning' derives in large part from carefully tailoring the hypothesis space so that the learner finds the right patterns with reasonable amounts of data and effort. Task-specific learning modules are familiar in biology (*30*) from examples such as the specialized neural mechanisms for birdsong learning (*31*) and the heritable variation in learning performance for foraging in honey bees (*32*).

Recent successful computational models of phonology (*33–35*) and syntax (*36–38*) demonstrate the utility of modular, domain-specific approaches to language learning. These learners

are successful because they look for different kinds of generalizations in words than in sentences.

Researchers who advocate a single general-purpose, domain-general learning model (*39*) must not only present a single learner capable of learning phonological patterns from words and syntactic patterns from sentences (to our knowledge no such demonstration exists), but must also explain the complexity differential. On the face of it, such models predict that sentence-like generalizations are possible within words and vice versa, contrary to the predictions of modular language learners. Evidence supporting this prediction is in principle possible to obtain. Artificial language-learning experiments (*40,41*) can determine whether people make generalizations of the same formal character in words and in sentences.

The hypothesis that human language learners consist of word-specific and sentence-specific learning modules currently offers the best explanation of the complexity differential observed above. The debate between domain-specific and domain-general models of human learning is not likely to be settled soon, and will certainly require continued experimentation and collaboration between computational learning theorists, experimental psychologists and neuroscientists. The computational complexity differential presents a new challenge to this old debate, one readily met by domain-specific learning.

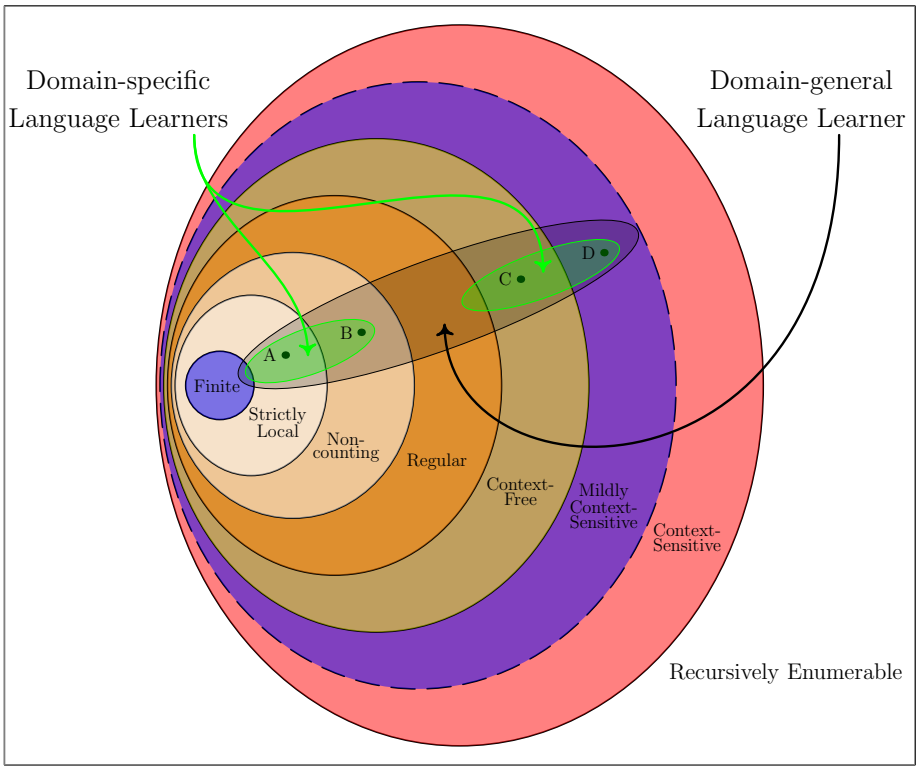

Figure 1: The Subregular and Chomsky Hierarchies (*10, 11*). A is a pattern of legal consonant clusters in English (*33*). B is the consonant harmony pattern of Samala (*21*). C is English nested embedding (*8*). D is Swiss German (*14*).

References and Notes

1. C. R. Gallistel, *The Cognitive Neurosciences*, M. Gazzaniga, ed. (MIT Press, 1999), pp. 1179–1191, second edn.

2. J. Spencer, M. S. C. Thomas, J. L. McClelland, *Toward a new unified theory of development: Connectionism and dynamical systems theory re-considered* (Oxford University Press, 2009).

3. M. A. Liberman, *Speech: A Special Code* (MIT Press, 1996).

4. E. Sapir, *Language* **1**, 37 (1925).

5. M. Halle, *Linguistic Theory and Psychological Reality* (The MIT Press, 1978).

6. N. Chomsky, *Syntactic Structures* (Mouton & Co., Printers, The Hague, 1957).

7. M. A. Nowak, N. L. Komarova, P. Niyogi, *Science* **291**, 114 (2001).

8. N. Chomsky, *IRE Transactions on Information Theory* (1956). IT-2.

9. W. M.S., *Nature* **309**, 118 (1984).

10. R. McNaughton, S. Papert, *Counter-Free Automata* (MIT Press, 1971).

11. N. Chomsky, *Information and Control* **2**, 137 (1956).

12. E. Charniak, *Statistical Language Learning* (MIT Press, 1996).

13. A. K. Joshi, *Natural Language Parsing*, D. Dowty, L. Karttunen, A. Zwicky, eds. (Cambridge University Press, 1985), pp. 206–250.

14. S. Shieber, *Linguistics and Philosophy* **8**, 333 (1985).

15. C. D. Johnson, *Formal Aspects of Phonological Description* (The Hague: Mouton, 1972).

16. R. Kaplan, M. Kay, *Computational Linguistics* **20**, 331 (1994).

17. M. D. Hauser, N. Chomsky, W. T. Fitch, *Science* **298**, 1569 (2002).

18. S. Rose, R. Walker, *Language* **80**, 475 (2004).

19. C. Ringen, *Vowel Harmony: Theoretical Implications* (Garland Publishing, Inc., 1988).

20. W. Poser, *The Structure of Phonological Representations*, H. van der Hulst, N. Smith, eds. (Dordrecht: Foris, 1982), pp. 121–158.

21. R. Applegate, *Samala-English dictionary : a guide to the Samala language of the Ineseño Chumash People* (Santa Ynez Band of Chumash Indians, 2007).

22. A. Gafos, *The Articulatory Basis of Locality in Phonology* (New York: Garland, 1999).

23. M. Tatham, K. Morton, *Speech Production and Perception* (Palgrave Macmillan, Basingstoke, 2006).

24. S. Öhman, *Journal of the Acoustic Society of America* **39**, 151 (1966).

25. E. Sober, *Evidence and Evolution* (Cambridge University Press, Cambridge, 2008).

26. M. A. Nowak, N. L. Komarova, P. Niyogi, *Nature* **417**, 611 (2002).

27. V. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).

28. S. Jain, D. Osherson, J. S. Royer, A. Sharma, *Systems That Learn: An Introduction to Learning Theory (Learning, Development and Conceptual Change)* (The MIT Press, 1999), second edn.

29. A. N. Meltzoff, P. K. Kuhl, J. Movellan, T. J. Sejnowski, *Science* **325**, 284 (2009).

30. C. R. Gallistel, *The Organization of Learning* (MIT Press, Cambridge, MA, 1990).

31. R. Mooney, *Current Opinion in Neurobiology* **19**, 654 (2009).

32. J. S. Latshaw, B. H. Smith, *Behavioral Ecology and Sociobiology* **58**, 200 (2005).

33. B. Hayes, C. Wilson, *Linguistic Inquiry* **39**, 379 (2008).

34. A. Albright, *Phonology* **26**, 9 (2009).

35. J. Heinz, *Phonology* **26**, 303 (2009).

36. A. Clark, R. Eyraud, *Journal of Machine Learning Research* **8**, 1725 (2007).

37. L. Becerra-Bonache, J. Case, S. Jain, F. Stephan, *19th International Conference on Algorithmic Learning Theory (ALT'08)* (Springer, 2008), vol. 5254, pp. 359–373. Expanded journal version accepted for the associated Special Issue of *TCS*, 2009.

38. R. Yoshinaka, *20th International Conference on Algorithmic Learning Theory (ALT'09)* (2009), vol. 5809 of *Lecture Notes in Artificial Intelligence*.

39. C. Kemp, J. B. Tenenbaum, *Proceedings of the National Academy of Sciences* **105**, 10687 (2008).

40. G. F. Marcus, S. Vijayan, S. B. Rao, P. Vishton, *Science* **283**, 77 (1999).

41. I. Berent, T. Lennertz, J. Jun, M. A. Moreno, P. Smolensky, *Proceedings of the National Academy of Sciences* **105**, 5321 (2008).

42. C. Kisseberth, *Linguistic Inquiry* **1**, 291 (1970).

43. E.-D. Cook, *A Sarcee Grammar* (University of British Columbia Press, 1984).

44. This work was supported by a University of Delaware Research Fund grant (to J.H.) and by NIH NIDCD grant (7R01DC005660-07) (to W.I.).