Authors: Heidi R. Waterfall and Shimon Edelman

Abstract words: 60
Text words: 967
References words: 450
Total words: 1477

# The Neglected Universals: Learnability Constraints and Discourse Cues

Heidi Waterfall
Dept. of Psychology, Cornell University
Ithaca, NY 14853, USA
and
Dept. of Psychology, University of Chicago
Chicago, IL 60637, USA
heidi.waterfall@gmail.com

Shimon Edelman
Dept. of Psychology, Cornell University
Ithaca, NY 14853, USA
and
Dept. of Brain and Cognitive Engineering, Korea University
Anam-dong, Seongbuk-gu, Seoul 136-713, South Korea
se37@cornell.edu
`http://kybele.psych.cornell.edu/~edelman`

**Abstract.** Converging findings from English, Mandarin, and other languages suggest that observed "universals" may be algorithmic. First, computational principles behind recently developed algorithms that acquire productive constructions from raw texts or transcribed child-directed speech impose family resemblance on learnable languages. Second, child-directed speech is particularly rich in statistical (and social) cues that facilitate learning of certain types of structures.

Having surveyed a wide range of posited universals and found them wanting, Evans and Levinson (E&L) propose instead that the "common patterns" observed in the organization of human languages are due to cognitive constraints and cultural factors. We offer empirical evidence in support of both these ideas.

One kind of common pattern is readily apparent in the six examples of child-directed speech in Figure 1, in each of which partial matches between successive utterances serve to highlight the structural regularities of the underlying language. Two universal principles that allow such regularities to be learned can be traced to the work of Zellig Harris (1946; 1991). First, the discovery of language structure, from morphemes to phrases, can proceed by cross-utterance alignment and comparison (Harris, 1946; Edelman and Waterfall, 2007). Second, the fundamental task in describing a language is to state the departures from equiprobability in its sound- and word-sequences (Harris, 1991, p.32; cf. Goldsmith, 2007).

These principles are precisely those used by the only two unsupervised algorithms capable of learning productive construction grammars from large-scale raw corpus data, ADIOS (Solan et al., 2005) and ConText (Waterfall et al., 2009). Both algorithms bootstrap from completely unsegmented text to words and to phrase structure by recursively identifying candidate constructions in patterns of partial alignment between utterances in the training corpus. Furthermore, in both algorithms, candidate structures must pass a statistical significance test before they join the growing grammar and the learning resumes (the algorithms differ in the way they represent corpus data and in the kinds of significance tests they impose).

The new algorithms exhibited hitherto unrivaled — albeit still very far from perfect — capacity for language learning, as measured by (1) precision, or acceptability of novel generated utterances, (2) recall, or coverage of withheld test corpus, (3) perplexity, or average uncertainty about the next lexical element in test utterances, and (4) performance in certain comprehension-related tasks (Edelman et al., 2004, 2005; Solan et al., 2005; Edelman and Solan, 2009). They have been tested, to varying extents, in English, French, Hebrew, Mandarin, and Spanish, to name but a few languages. The learning algorithms proved particularly effective when applied to raw transcribed child-directed speech (MacWhinney, 2000), achieving precision of 54% and 63% in Mandarin and English, and recall of about 30% in both languages (Solan et al., 2003; Brodsky et al., 2007).

To the extent that human learners rely on the same principles of aligning and comparing potentially relatable utterances, one may put these principles forward as the source of part of speech, phrase structure, and other structural "universals." In other words, certain forms may be common across languages because they are easier to learn, given the algorithmic constraints on the learner.[1]

Language acquisition becomes easier not only when linguistic forms match the algorithmic capabilities of the learner, but also when the learner's social environment is structured in various helpful ways. One possibility here is for mature speakers to embed structural cues in child-directed speech (CDS). Indeed, a growing body of evidence suggests that language acquisition is made easier than it would have been otherwise because of the way CDS is shaped by caregivers during their interaction with children.[2] One seemingly universal property of CDS is the prevalence of variation sets (Hoff-Ginsberg, 1990; Küntay and Slobin, 1996; Waterfall, 2006, 2009) — partial alignment among phrases uttered in temporal proximity, of the kind illustrated in Figure 1. The proportion of CDS utterances contained in variation sets is surprisingly constant across languages: 22% in Mandarin, 20% in Turkish, and 25% in English (when variation sets are defined by requiring consecutive caregiver utterances to have in common at least two lexical items in the same order; cf. Küntay and Slobin, 1996; this proportion grows to about 50% if a gap of two utterances is allowed between the partially matching ones). Furthermore, the lexical items (types) on which CDS utterances are aligned constitute a significant proportion of the corpus vocabulary, ranging from 9% in Mandarin to 32% in English.

Crucially, the nouns and verbs in variation sets in CDS were shown to be related to children's verb and noun use at the same observation, as well as to their production of verbs, pronouns, and subcategorization frames four months later (Hoff-Ginsberg, 1990; Waterfall, 2006, 2009). Moreover, experiments involving artificial language learning highlighted the causal role of variation sets: adults exposed to input that contained variation sets performed better in word segmentation and phrase boundary judgment tasks than a control group who heard the same utterances in a scrambled order, which had no variation sets (Onnis et al., 2008).

The convergence of the three lines of evidence mentioned above — the ubiquity of variation sets in

[1] Language may also be expected to evolve in the direction of a better fit between its structure and the learners' abilities (Christiansen and Chater, 2008).

[2] Social cues complement and reinforce structural ones in this context (Goldstein and Schwade, 2008).

child-directed speech in widely different languages, their proven effectiveness in facilitating acquisition, and the algorithmic revival of the principles of acquisition intuited by Harris — supports E&L's proposal of the origin of observed universals. More research is needed to integrate the computational framework outlined here with models of social interaction during acquisition and with neurobiological constraints on learning that undoubtedly contribute to the emergence of cognitive/cultural language universals.

## References

Brodsky, P., Waterfall, H. R., and Edelman, S. (2007). Characterizing Motherese: on the computational structure of child-directed language. In McNamara, D. S. and Trafton, J. G., editors, *Proceedings of the 29th Cognitive Science Society Conference*, Austin, TX. Cognitive Science Society.

Christiansen, M. H. and Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31:489–509.

Edelman, S. and Solan, Z. (2009). Translation using an automatically inferred structured language model. Submitted.

Edelman, S., Solan, Z., Horn, D., and Ruppin, E. (2004). Bridging computational, formal and psycholinguistic approaches to language. In *Proc. of the 26th Conference of the Cognitive Science Society*, Chicago, IL.

Edelman, S., Solan, Z., Horn, D., and Ruppin, E. (2005). Learning syntactic constructions from raw corpora. In *Proc. 29th annual Boston University Conference on Language Development*, Somerville, MA. Cascadilla Press.

Edelman, S. and Waterfall, H. R. (2007). Behavioral and computational aspects of language and its acquisition. *Physics of Life Reviews*, 4:253–277.

Goldsmith, J. A. (2007). Towards a new empiricism. In de Carvalho, J. B., editor, *Recherches linguistiques à Vincennes*, volume 36.

Goldstein, M. H. and Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, 19:515–523.

Harris, Z. S. (1946). From morpheme to utterance. *Language*, 22:161–183.

Harris, Z. S. (1991). *A theory of language and information*. Clarendon Press, Oxford.

Hoff-Ginsberg, E. (1990). Maternal speech and the child's development of syntax: A further look. *Journal of Child Language*, 17:85–99.

Küntay, A. and Slobin, D. (1996). Listening to a Turkish mother: Some puzzles for acquisition. In Slobin, D. and Gerhardt, J., editors, *Social interaction, social context, and language: Essays in honor of Susan Ervin-Tripp*, pages 265–286. Lawrence Erlbaum Associates, Hillsdale, NJ.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Erlbaum, Mahwah, NJ. Volume 1: Transcription format and programs. Volume 2: The Database.

Onnis, L., Waterfall, H. R., and Edelman, S. (2008). Learn locally, act globally: Learning language from variation set cues. *Cognition*, 109:423–430.

Solan, Z., Horn, D., Ruppin, E., and Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Science*, 102:11629–11634.

Solan, Z., Ruppin, E., Horn, D., and Edelman, S. (2003). Unsupervised efficient learning and representation of language structure. In Alterman, R. and Kirsh, D., editors, *Proc. 25th Conference of the Cognitive Science Society*, Hillsdale, NJ. Erlbaum.

Waterfall, H. R. (2006). *A little change is a good thing: Feature theory, language acquisition and variation sets*. PhD thesis, University of Chicago.

Waterfall, H. R. (2009). A little change is a good thing: The relation of variation sets to children's noun, verb and verb-frame development. Submitted.

Waterfall, H. R., Sandbank, B., Onnis, L., and Edelman, S. (2009). An empirical generative framework for computational modeling of language acquisition. Submitted.

English:
those are checkers
two checkers yes
play checkers

Italian:
dove sono
dove sono i coniglietti

Hebrew:
מה לא
מה את לא רוצה
את רוצה לספר

Korean:
제일 이뻐
누가 제일 이뻐
지원이 제일 이뻐 맞어

Mandarin:
这 是 什么 呀
哎呀 是 什么 呀

Russian:
вот твой папа не хочет с тобой остаться
как не хочет хочет
хочет папа хочет

Figure 1: Examples of child-directed speech in six languages. It is not necessary to be able to read, let alone understand, any of these languages to identify the most prominent structural characteristics common to these examples (see text for a hint). These characteristics should, therefore, be readily apparent to a prelinguistic baby, which is indeed the case, as the evidence we mention suggests. All the examples are from CHILDES corpora (MacWhinney, 2000).