Published in final edited form as:
Dev Sci. 2009 April ; 12(3): 409–411. doi:10.1111/j.1467-7687.2009.00827.x.

The learner as statistician: three principles of computational success in language acquisition

Melanie Soderstrom[1], **Erin Conwell**[2], **Naomi Feldman**[3], **James Morgan**[3]

[1]Department of Psychology, University of Manitoba, Winnipeg, Canada

[2]Department of Psychology, Harvard University, Cambridge, USA

[3]Department of Cognitive and Linguistic Sciences, Brown University, Providence, USA

Statistical learning is the new paradigm of language acquisition. A perusal of recent conference programs or journal contents reveals much work advocating – or criticizing – statistical learning. Language acquisition will continue to benefit from a variety of theories and methods, but, as the articles in this issue exemplify, statistical learning has progressed from being a minor player towards a central role. To ensure a lasting impact, statistical approaches must now move from piecemeal demonstrations towards a general theory of language learning.

Statistical learning stands in contrast to the predominant paradigm that it succeeded. The principles and parameters approach (Chomsky, 1981) assumed rich innate endowment, limited processing abilities and impoverished input, whereas statistical learning assumes that input is rich and that learners possess sufficient computational sophistication to extract relevant linguistic patterns. Statistical learning models are attractive because in principle they recruit powerful, task-general machinery to solve difficult problems of language acquisition. Furthermore, behavioural findings with both adults and infants suggest that humans use statistical learning in language-like tasks (Gómez, 2002; Goodsitt, Morgan & Kuhl, 1992; Maye, Werker & Gerken, 2002; Saffran, Aslin & Newport, 1996, Saffran, Newport & Aslin, 1996, inter alia).

However, extant models have been hand-crafted for particular problems, selecting relevant properties of input and learning mechanisms to arrive at predetermined output structures. Although these models may provide proofs-in-principle on possible ways to solve language learning problems, they do not illuminate how the learner knows how to select and organize the input appropriately for any particular task, what the most appropriate output representations might be, or how the learner chooses specific statistical analyses to pursue. Different aspects of language will require different computational solutions – hence learners need some way of selecting *particular* computations for *particular* problems. Assuming that learners attend to certain properties of input and have a repertoire of computational skills, how do they match the latter to the former to obtain appropriate outcomes?

Address for correspondence: Melanie Soderstrom, Department of Psychology, P404 Duff Roblin Building, 190 Dysart Road, University of Manitoba, Winnipeg, MB, Canada R3T 2N2; M_Soderstrom@umanitoba.ca.

Eschewing teleology, we begin with the hypothesis that infants pursue multiple analyses of many input properties, letting a thousand flowers bloom. The problem then is how to winnow these: what evaluation metric can learners use? Other things being equal, analyses that provide suitable descriptions of new input and are generalizable should be preferred, as should analyses that provide relatively compact descriptions (for example as can be implemented in a minimum description length or a Bayesian framework; see Brent, 1999; Jeffreys & Berger, 1992). We suggest three additional principles that may help learners to decide which analyses to retain: heterogeneity of variance, convergence of cues, and bootstrapping. The four papers in this special section illustrate how these principles may apply.

Heterogeneity of variance

A critical task for the language learner is to determine the units over which an analysis is made. Language is characterized by patterns of opposition and alternation: greater variability coincides with transitions between units. Learners can exploit these patterns to locate relevant units as well as the boundaries between units. Therefore, input representations that generate peaks and valleys of variance are likely to prove fruitful. Conversely, homogeneity of variance may indicate to the learner that a particular analysis of the input should be abandoned.

The usefulness of heterogeneity of variance is exemplified by distributional learning hypotheses, investigated in articles in this section as a means of finding categories. Boundaries between categories may be characterized by high variance; low variance indicates that items are members of the same category. For example, although in general there is considerable variance among speech sounds, most speech sounds are contained in several distinct low-variance regions of acoustic space. McMurray, Aslin and Toscano (this issue) hypothesize that learners can use this pattern of variability to recognize that speech sounds can be grouped into categories. They present an incremental learning algorithm that capitalizes on these patterns of variability to discover the locations of individual categories, placing the centres of the Gaussian categories in regions of low variability. In line with the principle of heterogeneity of variance, their algorithm performs best when it hypothesizes a ‘sparse’ solution that highlights the low variability within speech sound categories. The work by Christiansen, Onnis and Hockema (this issue) also uses high between-category variation and low within-category variation, in this case to categorize words as noun or verb. In particular, they find that the phonemes at word edges are somewhat predictive of lexical category. Again, this demonstrates that learners would do well to look for points of changing variance in the data. Where variability is low, words can be classified together, and where variability is high, words should be treated as members of different classes.

The notion of heterogeneity of variance plays out in a different way in the work on transitional probabilities as a means of finding boundaries between linguistic units within larger sequences. Christiansen *et al.* show that the majority of phoneme transitions in English can be divided into two classes with unequal variance: those that appear word-medially and those that appear across word boundaries. Such heterogeneity of variance may serve as a clue to the learner that transitional probabilities between phonemes will provide useful information for segmentation.

A third role for heterogeneity of variance is exemplified by the frequent-frames model, discussed by Chemla, Mintz, Bernal and Christophe (this issue). The invariance of frequently occurring [A x B] frames contrasts with the highly variable *x* elements between, as well as with elements found before and after. Furthermore, the importance of the discontinuity of framing elements may be viewed from the perspective of this principle. Comparing the discontinuous frame, [A x B], with prefixed [A B x] and suffixed [x A B] contexts, note that the frame context contains an additional source of variance – the transitional probabilities after *A* and before *B*, whereas the prefixed and suffixed contexts only have before *A* and after *B*. This additional piece of information may be crucial to the learner detecting the frame as a reliable source of contextual information because it highlights the contrast in variability between the *x* element and the surrounding frame.

Convergence of cues

Important properties of language are often clued by multiple input cues. For example, sentential phrase structure may be cued by semantics (phrases are meaningful substructures), prosody (e.g. phrase-final lengthening), locations of function words (at leading or trailing edges of phrases), morphological concord (holding within phrasal domains), and patterns of substitution and privileges of occurrence. Similarly, grammatical categories may be cued by distributional patterns, phonotactic and prosodic properties, semantic reference, and so forth. Converging analyses based on a variety of input properties are mutually reinforcing; individual analyses pointing in unique directions should be treated with scepticism. In line with this, Hollich and Prince (this issue) show that audiovisual synchronization can help explain infants' looking behaviour towards audiovisual stimuli. Infants are more attentive when two dimensions are correlated with each other, demonstrating that they attend to – and perhaps seek out – correlations among cues.

In the domain of syntax, cue convergence could provide a partial explanation for how the incomplete categories created by a frequent-frames analysis are combined into grammatical categories. If Christiansen *et al.*'s (this issue) results are taken in combination with Chemla *et al.*'s (this issue) findings, words that appear in similar frames may also share certain phonological properties. The fact that these cues both point to the same conclusion would indicate to the learner that they are relevant to categorization.

Bootstrapping

Learning about syntax requires analysis of patterns of co-occurrence of words. Words, however, are not given in the input, but rather must be discovered (in part) through analysis of segmented portions of the speech stream. Thus, the *output* of one analysis often serves as the *input* of another. This principle straddles analyses, concerning how one co-ordinates with another. Such analyses are often, although not always, hierarchically related. An analysis of input in one instance may automatically provide the units for analysis in another, in an upward spiral of linguistic ability. Analyses that successfully feed into additional analyses should be privileged by the learner.

Of course, this does not mean that an analysis must be perfected at one level to inform the next. Christiansen *et al.* (this issue) nicely illustrate that even noisy or incomplete results from one language learning process can be useful for another. Although their segmentation results are imperfect, these segmentations nevertheless provide a sufficient basis for preliminary lexical categorization. This suggests that language learners can use incomplete or even partially inaccurate results to begin another language learning process, allowing learning to proceed in parallel, with revised or updated outputs of one system continuously being used for other processes. Such a mechanism is potentially valuable for explaining the rapid rate of early language learning.

Conclusion

The principles that we have proposed here are among many that might guide statistical learning. Which are operative in children's language acquisition is, of course, an empirical question. The time has come, however, for statistical learning to progress to the next stage, to move from isolated demonstrations, however compelling, towards a comprehensive theory. Consideration of the sorts of principles we discuss here will, we believe, be instrumental in taking this next step.

References

Brent MR (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. Machine Learning, 34, 71–105.

Chemla E, Mintz TH, Bernal S, & Christophe A (2009). Categorizing words using 'frequent fames': what cross-linguistic analyses reveal about distributional acquisition strategies. Developmental Science, 12 (3), 397–407.

Chomsky N (1981). Lectures on Government and Binding. Berlin: Mouton De Gruyter.

Christiansen MH, Onnis L, & Hockema SA (2009). The secret is in the sound: from unsegmented speech to lexical categories. Developmental Science, 12 (3), 388–396. [PubMed: 19371361]

Gómez RL (2002). Variability and detection of invariant structure. Psychological Science, 13, 431–436. [PubMed: 12219809]

Goodsitt J, Morgan JL, & Kuhl PK (1993). Perceptual strategies in prelingual speech segmentation. Journal of Child Language, 20, 229–252. [PubMed: 8376468]

Hollich G, & Prince CG (2009). Comparing infants' preference for correlated audiovisual speech with signal-level computational models. Developmental Science, 12 (3), 379–387. [PubMed: 19371360]

Jeffreys WH, & Berger JO (1992). Ockham's razor and Bayesian analysis. American Scientist, 80, 64–72.

McMurray B, Aslin RN, & Toscano JC (2009). Statistical learning of phonetic categories: insights from a computational approach. Developmental Science, 12 (3), 369–378. [PubMed: 19371359]

Maye J, Werker JF, & Gerken L (2002). Infant sensitivity to distributional information can affect phonetic discrimination. Cognition, 82 (3), B101–B111. [PubMed: 11747867]

Saffran JR, Aslin RN, & Newport EL (1996). Statistical learning by 8-month old infants. Science, 274, 1926–1928. [PubMed: 8943209]

Saffran JR, Newport EL, & Aslin RN (1996). Word segmentation: the role of distributional cues. Journal of Memory and Language, 35, 606–621.