Statistics Learning and Universal Grammar: Modeling Word Segmentation

Timothy Gambell
59 Bishop Street
New Haven, CT 06511
USA
timothy.gambell@aya.yale.edu

Charles Yang
Department of Linguistics, Yale University
New Haven, CT 06511
USA
charles.yale.edu@yale.edu

Abstract

This paper describes a computational model of word segmentation and presents simulation results on realistic acquisition In particular, we explore the capacity and limitations of statistical learning mechanisms that have recently gained prominence in cognitive psychology and linguistics.

1 Introduction

Two facts about language learning are indisputable. First, only a human baby, but not her pet kitten, can learn a language. It is clear, then, that there must be some element in our biology that accounts for this unique ability. Chomsky's Universal Grammar (UG), an innate form of knowledge specific to language, is an account of what this ability is. This position gains support from formal learning theory [1-3], which sharpens the logical conclusion [4,5] that no (realistically efficient) learning is possible without priori restrictions on the learning space. Second, it is also clear that no matter how much of a head start the child has through UG, language *is* learned. Phonology, lexicon, and grammar, while governed by universal principles and constraints, do vary from language to language, and they must be learned on the basis of linguistic experience. In other words–indeed a truism–both endowment and learning contribute to language acquisition, the result of which is extremely sophisticated body of linguistic knowledge. Consequently, both must be taken in account, explicitly, in a theory of language acquisition [6,7].

Controversies arise when it comes to the relative contributions by innate knowledge and experience-based learning. Some researchers, in particular linguists, approach language acquisition by characterizing the scope and limits of innate principles of Universal Grammar that govern the world's language. Others, in particular psychologists, tend to emphasize the role of experience and the child's domain-general learning ability. Such division of research agenda understandably stems from the division of labor between endowment and learning–plainly, things that are built in needn't be learned, and things that can be garnered from experience needn't be built in.

The important paper of Saffran, Aslin, & Newport [8] on statistical learning (SL), suggests that children may be powerful learners after all. Very young infants can exploit transitional probabilities between syllables for the task of word segmentation, with only minimum exposure to an artificial language. Subsequent work has demonstrated SL in other domains including artificial grammar learning [9], music [10], vision [11], as well as in other species [12]. This then raises the possibility of learning as an alternative to the innate endowment of linguistic knowledge [13].

We believe that the computational modeling of psychological processes, with special attention to concrete mechanisms and quantitative evaluations, can play an important role in the endowment vs. learning debate. Linguists' investigations of UG are rarely developmental, even less so corpus-oriented. Developmental psychologists, by contrast, often stop at identifying *components* in a cognitive task [14], without an account of how such components work together in an algorithmic manner. On the other hand, if computation is to be of relevance to linguistics, psychology, and cognitive science in general, being merely computational will not suffice. A model must be psychological plausible, and ready to face its implications in the broad empirical contexts [7]. For example, how does it generalize to typologically different languages? How does the model's behavior compare with that of human language learners and processors?

In this article, we will present a simple computational model of word segmentation and some of its formal and developmental issues in child language acquisition. Specifically we show that SL using transitional probabilities cannot reliably segment words when scaled to a realistic setting (e.g., child-directed English). To be successful, it must be constrained by the knowledge of phonological structure. Indeed, the model reveals that SL may well be an artifact–an impressive one, nonetheless–that plays no role in actual word segmentation in human children.

2 Statistics does not Refute UG

It has been suggested [15, 8] that word segmentation from continuous speech may be achieved by using transitional probabilities (TP) between adjacent syllables A and B, where , TP(A→B) = P(AB)/P(A), with P(AB) being the frequency of B following A, and P(A) the total frequency of A. Word boundaries are postulated at local minima, where the TP is lower than its neighbors. For example, given sufficient amount of exposure to English, the learner may establish that, in the four-syllable sequence "prettybaby", TP(pre→tty) and TP(ba→by) are both higher than TP(tty→ba): a word boundary can be (correctly) postulated. It is remarkable that 8-month-old infants can extract three-syllable words in the continuous speech of an artificial language from only two minutes of exposure [8].

To be effective, a learning algorithm–indeed any algorithm–must have an appropriate representation of the relevant learning data. We thus need to be cautious about the interpretation of the success of SL, as the authors themselves note [16]. If anything, it seems that the findings strengthen, rather than weaken, the case for (innate) linguistic knowledge. A classic argument for innateness [4, 5, 17] comes from the fact that syntactic operations are defined over specific types of data structures–constituents and phrases–but not over, say, linear strings of words, or numerous other logical possibilities. While infants seem to keep track of statistical information, any conclusion drawn from such findings must presuppose children knowing what kind of statistical information to keep track of. After all, an infinite range of statistical correlations exists in the acoustic input: e.g., What is the probability of a syllable rhyming with the next? What is the probability of two adjacent vowels being both nasal? The fact that infants can use SL to segment syllable sequences at all entails that, at the minimum, they know the relevant unit of information over which correlative statistics is gathered: in this case, it is the syllables, rather than segments, or front vowels.

A host of questions then arises. First, How do they know so? It is quite possible that the primacy of syllables as the basic unit of speech is innately available, as suggested in neonate speech perception studies [18]? Second, where do the syllables come from? While the experiments in [8] used uniformly CV syllables, many languages, including English, make use of a far more diverse range of syllabic types. And then, syllabification of speech is far from trivial, which (most likely) involve both innate knowledge of phonological structures as well as discovering language-specific instantiations [14]. All these problems have to be solved before SL for word segmentation can take place.

3 The Model

To give a precise evaluation of SL in a realistic setting, we constructed a series of (embarrassingly simple) computational models tested on child-directed English.

The learning data consists of a random sample of child-directed English sentences from the CHILDES database [19] The words were then phonetically transcribed using the Carnegie Mellon Pronunciation Dictionary, and were then grouped into syllables. Spaces between words are removed; however, utterance breaks are available to the modeled learner. Altogether, there are 226,178 words, consisting of 263,660 syllables.

Implementing SL-based segmentation is straightforward. One first gathers pair-wise TPs from the training data, which are used to identify local minima and postulate word boundaries in the on-line processing of syllable sequences. Scoring is done for each utterance and then averaged. Viewed as an information retrieval problem, it is customary [20] to report both precision and recall of the performance.

The segmentation results using TP local minima are remarkably poor, even under the assumption that the learner has already syllabified the input perfectly. Precision is 41.6%, and recall is 23.3%; over half of the words extracted by the model are not actual English words, while close to 80% of actual words fail to be extracted. And it is straightforward why this is the case. In order for SL to be effective, a TP at an actual word boundary must be lower than its neighbors. Obviously, this condition cannot be met if the input is a sequence of monosyllabic words, for which a space must be postulated for every syllable; there are no local minima to speak of. While the pseudowords in [8] are uniformly three-syllables long, much of child-directed English consists of sequences of monosyllabic words: corpus statistics reveals that on average, a monosyllabic word is followed by another monosyllabic word 85% of time. As long as this is the case, SL cannot, in principle, work.

4 Statistics Needs UG

This is not to say that SL cannot be effective for word segmentation. Its application, must be constrained–like that of any learning algorithm however powerful–as suggested by formal learning theories [1-3]. The performance improves dramatically, in fact, if the learner is equipped with even a small amount of prior knowledge about phonological structures. Specifically, we assume, uncontroversially, that each word can have only one primary stress. (This would not work for functional words, however.) If the learner knows this, then it may limit the search for local minima only in the window between two syllables that both bear primary stress, e.g., between the two **a**'s in the sequence "l**a**nguage**a**cquisition". This assumption is plausible given that 7.5-month-old infants are sensitive to strong/weak prosodic distinctions [14]. When stress information suffices, no SL is employed, so "b**i**gb**a**dw**o**lf" breaks into three words for free. Once this simple principle is built in, the stress-delimited SL algorithm can achieve the precision of 73.5% and 71.2%, which compare favorably to the best performance reported in the literature [20]. (That work, however, uses an computationally prohibitive and psychological implausible algorithm that iteratively optimizes the entire lexicon.)

The computational models complement the experimental study that prosodic information takes priority over statistical information when both are available [21]. Yet again one needs to be cautious about the improved performance, and a number of unresolved issues need to be addressed by future work. It remains possible that SL is not used at all in actual word segmentation. Once the one-word-one-stress principle is built in, we may consider a model that does not use any statistics, hence avoiding the computational cost that is likely to be considerable. (While we don't know how infants keep track of TPs, there are clearly quite some work to do. Syllables in English number in the thousands; now take the quadratic for the potential number of pair-wise TPs.) It simply stores previously extracted words in the memory to bootstrap new words. Young children's familiar segmentation errors–"I was have" from be-have, "hiccing up" from hicc-up, "two dults", from a-dult–suggest that this process does take place. Moreover, there is evidence that 8-month-old infants can store familiar sounds in the memory [22]. And finally, there are plenty of single-word utterances–up to 10% [23]–that give many words for free. The implementation of a purely symbolic learner that recycles known words yields even better performance: a precision of 81.5% and recall of 90.1%.

5 Conclusion

Further work, both experimental and computational, will need to address a few pressing questions, in order to gain a better assessment of the relative contribution of SL and UG to language acquisition. These include, more pertinent to the problem of word segmentation:

- Can statistical learning be used in the acquisition of language-specific phonotactics, a prerequisite to syllabification and a prelude to word segmentation?
- Given that prosodic constraints are critical for the success of SL in word segmentation, future work needs to quantify the availability of stress information in spoken corpora.
- Can further experiments, carried over realistic linguistic input, further tease apart the multiple strategies used in word segmentation [14]? What are the psychological mechanisms (algorithms) that integrate these strategies?
- How does word segmentation, statistical or otherwise, work for agglutinative (e.g., Turkish) and polysynthetic languages (e.g. Mohawk), where the division between words, morphology, and syntax is quite different from more clear-cut cases like English?

Computational modeling can make explicit the balance between statistics and UG, and are in the same vein as the recent findings [24] on when/where SL is effective/possible. UG can help SL by providing specific constraints on its application, and modeling may raise new questions for further experimental studies. In related work [6,7], we have augmented traditional theories of UG–derivational phonology, and the Principles and Parameters framework–with a component of statistical learning, with novel and desirable consequences. Yet in all cases, statistical learning, while perhaps domain-general, is constrained by what appears to be innate and domain-specific knowledge of linguistic structures, such that learning can operate on specific aspects of the input evidence

References

1. Gold, E. M. (1967). Language identification in the limit. *Information and Control*, 10:447-74.

2. Valiant, L. (1984). A theory of the learnable. *Communication of the ACM*. 1134-1142.

3. Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Berlin: Springer.

4. Chomsky, N. (1959). Review of Verbal Behavior by B.F. Skinner. *Language*, 35, 26-57.

5. Chomsky, N. (1975). *Reflections on Language*. New York: Pantheon.

6. Yang, C. D. (1999). A selectionist theory of language development. In *Proceedings of 37th Meeting of the Association for Computational Linguistics*. Maryland, MD. 431-5.

7. Yang, C. D. (2002). *Knowledge and Learning in Natural Language*. Oxford: Oxford University Press.

8. Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month old infants. *Science,* 274, 1926-1928.

9. Gomez, R.L., & Gerken, L.A. (1999). Artificial grammar learning by one-year-olds leads to specific and abstract knowledge. *Cognition*, 70, 109-135.

10. Saffran, J.R., Johnson, E.K., Aslin R.N. & Newport, E.L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70, 27-52.

11. Fiser, J., & Aslin, R.N. (2002). Statistical learning of new visual feature combinations by infants. *PNAS*, 99, 15822-6.

12. Hauser, M., Newport, E.L., & Aslin, R.N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, 78, B41-B52.

13. Bates, E., & Elman, J. (1996). Learning rediscovered. *Science*, 274, 1849-50.

14. Jusczyk, P.W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Sciences*, 3, 323-8.

15. Chomsky, N. (1955/1975). *The Logical Structure of Linguistic Theory*. Manuscript, Harvard University and Massachusetts Institute of Technology. Published in 1975 by New York: Plenum.

16. Saffran, J.R., Aslin, R.N., & Newport, E.L. (1997). Letters. *Science*, 276, 1177-1181

17. Crain, S., & Nakayama, M. (1987). Structure dependency in grammar formation. *Language*, 63:522-543.

18. Bijeljiac-Babic, R., Bertoncini, J., & Mehler, J. (1993). How do four-day-old infants categorize multisyllabic utterances. *Developmental psychology*, 29, 711-21.

19. MacWhinney, B. (1995). *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale: Lawrence Erlbaum.

20. Brent, M. (1999). Speech segmentation and word discovery: a computational perspective. *Trends in Cognitive Science*, 3, 294-301.

21. Johnson, E.K. & Jusczyk, P.W. (2001) Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language,* 44, 1-20.

22. Jusczyk, P. W., & Hohne, E. A. (1997). Infants' memory for spoken words. *Science,* 277, 1984-6.

23. Brent, M.R., & Siskind, J.M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, 81, B33-44.

24. Newport, E.L., & Aslin, R.N. (2004). Learning at a distance: I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, 48, 127-62.