Psych 229: Language Acquisition

Lecture 5

Statistics & Words

Saffran, Aslin, & Newport 1996

Depending on the developmental status and the task facing a particular organism, both experience-independent and experience-dependent mechanisms may be involved in the extraction of information and the control of behavior.

In the domain of language acquisition, two facts have supported the interpretation that experience-independent mechanisms are both necessary and dominant. First, highly complex forms of language production develop extremely rapidly (3).

Second, the language input available to the young child is both incomplete and sparsely represented compared to the child's eventual linguistic abilities (4). Thus, most theories of language acquisition have emphasized the critical role played by experience-independent internal structures over the role of experience-dependent factors (5).

It is undeniable that experience-dependent mechanisms are also required for the acquisition of language. Many aspects of a particular natural language must be acquired from listening experience. For example, acquiring the specific words and phonological structure of a language requires exposure to a significant corpus of language input. Moreover, long before infants begin to produce their native language, they acquire information about its sound properties.

What does experience-independent mean (as opposed to experience-dependent)?

the task at hand

One task faced by all language learners is the segmentation of fluent speech into words. This process is particularly difficult because word boundaries in fluent speech are marked inconsistently by discrete acoustic events such as pauses.

Saffran, Aslin, & Newport 1996

One important source of information that can, in principle, define word boundaries in any natural language is the statistical information contained in sequences of sounds. Over a corpus of speech there are measurable statistical regularities that distinguish recurring sound sequences that comprise words from the more accidental sound sequences that occur across word boundaries (11).

transitional probability = conditional probability

$$Y|X = \frac{\text{frequency of XY}}{\text{frequency of X}}$$

Within a language, the transitional probability from one sound to the next will generally be highest when the two sounds follow one another within a word, whereas transitional probabilities spanning a word boundary will be relatively low (12).

For example, given the sound sequence pretty#baby, the transitional probability from *pre* to *ty* is greater than the transitional probability from *ty* to *ba*.

We asked whether 8-month-old infants can extract information about word boundaries solely on the basis of the sequential statistics of concatenated speech.

We used the familiarization-preference procedure developed by Jusczyk and Aslin (9). In this procedure, infants are exposed to auditory material that serves as a potential learning experience. They are subsequently presented with two types of test stimuli: (i) items that were contained within the familiarization material and (ii) items that are highly similar but (by some critical criterion) were not contained within the familiarization material. During a series of test trials that immediately follows familiarization, infants control the duration of each test trial by their sustained visual fixation on a blinking light (14). If infants have extracted the crucial information about the familiarization items, they may show differential durations of fixation (listening) during the two types of test trials (15).

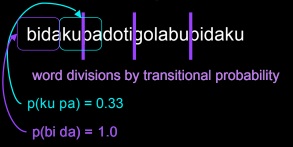
Why do they need the light, too?

Saffran, Aslin, & Newport 1996

In our first experiment, 24 8-month-old infants from an American-English language environment were familiarized with 2 min of a continuous speech stream consisting of four three-syllable nonsense words (hereafter, "words") repeated in random order (16). The speech stream was generated by a speech synthesizer in a monotone female voice at a rate of 270 syllables per minute (180 words in total).

A sample of the speech stream is the orthographic string *bidakupadotigolabubidaku*. . . . The only cues to word boundaries were the transitional probabilities between syllable pairs, which were higher within words (1.0 in all cases, for example, *bida*) than between words (0.33 in all cases, for example, *kupa*).

To assess learning, each infant was presented with repetitions of one of four three-syllable strings on each test trial. Two of these three-syllable strings were "words" from the artificial language presented during familiarization, and two were three-syllable "nonwords" that contained the same syllables heard during familiarization but not in the order in which they appeared as words (17).



bidaku|padoti|golabu|bidaku

word divisions by transitional probability

p(ku pa) = 0.33

p(bi da) = 1.0

Another 24 8-month-old infants from an American-English language environment were familiarized with 2 min of a continuous speech stream consisting of three-syllable nonsense words similar in structure to the artificial language used in our first experiment (19). This time, however, the test items for each infant consisted of two words and two "part-words." The part-words were created by joining the final syllable of a word to the first two syllables of another word.

Saffran, Aslin, & Newport 1996

Despite the difficulty of this word versus part-word discrimination, infants showed a significant test-trial discrimination between the word and part-word stimuli (*21*), with longer listening times for part-words (Table 1).

infants succeeded in learning and remembering particular groupings of three-syllable strings—those strings containing higher transitional probabilities surrounded by lower transitional probabilities.

Table 1. Mean time spent listening to the familiar and novel stimuli for experiment 1 (words versus nonwords) and experiment 2 (words versus part-words) and significance tests comparing the listening times.

Experiment	Mean listening times (s)		Matched-pairs *t* test
	Familiar items	Novel items	
1	7.97 (SE = 0.41)	8.85 (SE = 0.45)	$t(23) = 2.3, P < 0.04$
2	6.77 (SE = 0.44)	7.60 (SE = 0.42)	$t(23) = 2.4, P < 0.03$

novelty preference

The infants' performance in these studies is particularly impressive given the impoverished nature of the familiarization speech stream, which contained no pauses, intonational patterns, or any other cues that, in normal speech, probabilistically supplement the sequential statistics inherent in the structure of words.

Saffran, Aslin, & Newport 1996

Although experience with speech in the real world is unlikely to be as concentrated as it was in these studies, infants in more natural settings presumably benefit from other types of cues correlated with statistical information.

So this isn't about learning *only* from transitional probabilities…

It remains unclear whether the statistical learning we observed is indicative of a mechanism specific to language acquisition or of a general learning mechanism applicable to a broad range of distributional analyses of environmental input (22).

And this statistical tracking ability may apply only to language data…

In particular, some aspects of early development may turn out to be best characterized as resulting from innately biased statistical learning mechanisms rather than innate knowledge.

Innate knowledge in the form of biases on learning, rather than explicit knowledge.

Discussion Questions

How does statistical learning fit in with the idea of a mental grammar? What about with the idea of innateness? Can experience-independent mechanisms ensure learning by themselves in some situation?

Transitional probability: how does this fit into experience-dependent and experience-independent learning mechanisms?

Gambell & Yang 2006: Computational model of word segmentation

Survey of infant strategies (use at 8 months [before word meaning])

Possible strategy: learn from isolated words

- Data: 9% of mother-to-child speech is isolated words
- Problem: How does a child recognize an isolated word as such?
  - length won't work: "I-see" vs. "spaghetti"

Possible strategy: statistical properties like transitional probability between syllables

- word boundaries postulated at local minima
  - pre tty ba by   p(tty-->ba) < p(pre-->tty), p(ba-->by)

- Question: How well does this fare on real data sets (not artificial stimuli)?

Gambell & Yang 2006: Computational model of word segmentation

Survey of infant strategies (use at 8 months [before word meaning])

Possible strategy: Metrical segmentation strategy

- Children treat stressed syllable as beginning of word
- - 90% of English content words are stress-initial
- Problem: Stress systems differ from language to language
  - - the child would need to know that words are stress initial
  - …but to do that, the child needs words *first*

Possible strategy: phonotactic constraints (sequences of consonant clusters that go together, e.g. str vs. *stl in English); language-specific

- - Infants seem to know these by 9 months
- - posit boundary at improper sequence break: stl --> st l (first light)
- Problem: May just be syllable boundary (restless)

Gambell & Yang 2006: Computational model of word segmentation

Survey of infant strategies (use at 8 months [before word meaning])

Possible strategy: Memory

- Use previous stored words (sound forms, not meanings) to recognize new words
- - if child knows *new*, then can recognize *one* in *thatsanewone*
- Problem: Needs to know words before can use this

A good point: "It seems…only language-independent strategies can set word segmentation in motion before the establishment and application of language-specific strategies"

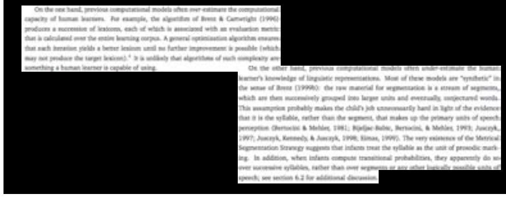
Gambell & Yang 2006: Computational model of word segmentation

Computational model goal

- psychologically plausible learning algorithm
- real data

Another good point: it's good if the information is in the data, but we also need to know how children could use it

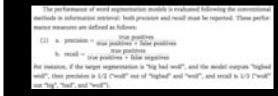
On psychological plausibility



On the one hand, previous computational models often over-estimate the computational capacity of human learners. For example, the algorithm of Brent & Cartwright (1996) produces a succession of lexicons, each of which is associated with an evaluation metric that is calculated over the entire learning corpus. A general optimization algorithm ensures that each iteration yields a better lexicon until no further improvement is possible (which may not produce the target lexicon).[4] It is unlikely that algorithms of such complexity are something a human learner is capable of using.

On the other hand, previous computational models often under-estimate the human learner's knowledge of linguistic representations. Most of these models are "synthetic" in the sense of Brent (1999b): the raw material for segmentation is a stream of segments, which are then successively grouped into larger units and eventually, conjectured words. This assumption probably makes the child's job unnecessarily hard in light of the evidence that it is the syllable, rather than the segment, that makes up the primary units of speech perception (Bertoncini & Mehler, 1981; Bijeljac-Babic, Bertoncini, & Mehler, 1993; Jusczyk, 1997; Jusczyk, Kennedy, & Jusczyk, 1998; Eimas, 1999). The very existence of the Metrical Segmentation Strategy suggests that infants treat the syllable as the unit of prosodic marking. In addition, when infants compute transitional probabilities, they apparently do so over successive syllables, rather than over segments or any other logically possible units of speech; see section 6.2 for additional discussion.

Gambell & Yang 2006: Computational model of word segmentation



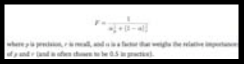
The performance of word segmentation models is evaluated following the conventional methods in information retrieval: both *precision* and *recall* must be reported. These performance measures are defined as follows:

(1) a. precision = $\frac{\text{true positives}}{\text{true positives} + \text{false positives}}$

b. recall = $\frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$

For instance, if the target segmentation is "big bad wolf", and the model outputs "bigbad wolf", then precision is 1/2 ("wolf" out of "bigbad" and "wolf"), and recall is 1/3 ("wolf" out "big", "bad", and "wolf").

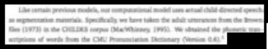
what to evaluate



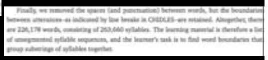
$$F = \frac{1}{\alpha\frac{1}{p} + (1-\alpha)\frac{1}{r}}$$

where p is precision, r is recall, and α is a factor that weighs the relative importance of p and r (and is often chosen to be 0.5 in practice).

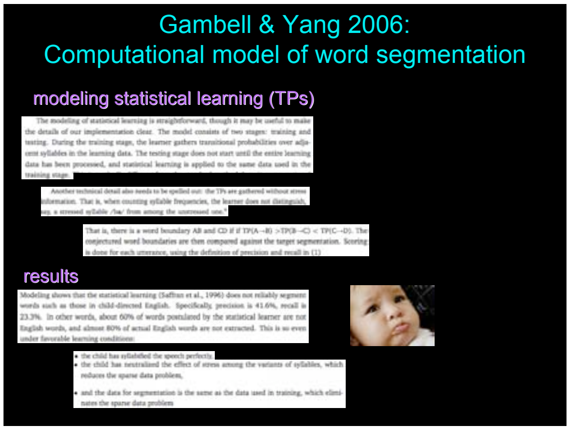
where to get the data



Like certain previous models, our computational model uses actual child-directed speech as segmentation materials. Specifically, we have taken the adult utterances from the Brown files (1973) in the CHILDES corpus (MacWhinney, 1995). We obtained the phonetic transcriptions of words from the CMU Pronunciation Dictionary (Version 0.6).[5]



Finally, we removed the spaces (and punctuation) between words, but the boundaries between utterances–as indicated by line breaks in CHILDES–are retained. Altogether, there are 226,178 words, consisting of 263,660 syllables. The learning material is therefore a list of unsegmented syllable sequences, and the learner's task is to find word boundaries that group substrings of syllables together.



Gambell & Yang 2006:
Computational model of word segmentation

modeling statistical learning (TPs)

The modeling of statistical learning is straightforward, though it may be useful to make the details of our implementation clear. The model consists of two stages: training and testing. During the training stage, the learner gathers transitional probabilities over adjacent syllables in the learning data. The testing stage does not start until the entire learning data has been processed, and statistical learning is applied to the same data used in the training stage.

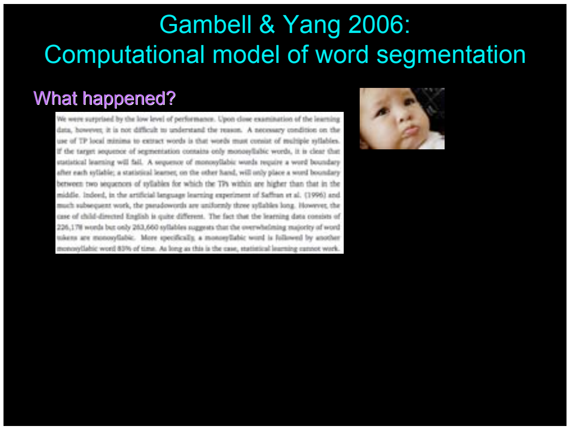
Another technical detail also needs to be spelled out: the TPs are gathered without stress information. That is, when counting syllable frequencies, the learner does not distinguish, say, a stressed syllable /ba/ from among the unstressed one.

That is, there is a word boundary AB and CD if if TP(A→B) >TP(B→C) < TP(C→D). The conjectured word boundaries are then compared against the target segmentation. Scoring is done for each utterance, using the definition of precision and recall in (1)

results

Modeling shows that the statistical learning (Saffran et al., 1996) does not reliably segment words such as those in child-directed English. Specifically, precision is 41.6%, recall is 23.3%. In other words, about 60% of words postulated by the statistical learner are not English words, and almost 80% of actual English words are not extracted. This is so even under favorable learning conditions:

- the child has syllabified the speech perfectly,
- the child has neutralized the effect of stress among the variants of syllables, which reduces the sparse data problem,
- and the data for segmentation is the same as the data used in training, which eliminates the sparse data problem



Gambell & Yang 2006:
Computational model of word segmentation

What happened?

We were surprised by the low level of performance. Upon close examination of the learning data, however, it is not difficult to understand the reason. A necessary condition on the use of TP local minima to extract words is that words must consist of multiple syllables. If the target sequence of segmentation contains only monosyllabic words, it is clear that statistical learning will fail. A sequence of monosyllabic words require a word boundary after each syllable; a statistical learner, on the other hand, will only place a word boundary between two sequences of syllables for which the TPs within are higher than that in the middle. Indeed, in the artificial language learning experiment of Saffran et al. (1996) and much subsequent work, the pseudowords are uniformly three syllables long. However, the case of child-directed English is quite different. The fact that the learning data consists of 226,178 words but only 263,660 syllables suggests that the overwhelming majority of word tokens are monosyllabic. More specifically, a monosyllabic word is followed by another monosyllabic word 85% of time. As long as this is the case, statistical learning cannot work.