Overview

1) Summary, Days 1-3
   a) What is a syllable: meta-linguistically and linguistically (and why both)
   b) Evidence:
      i) On-line evidence in French, but not in English perhaps due to Kahn-type
         resyllabification and/or stress as Cutler suggests
      ii) Off-line evidence in English for a different kind of syllable than on-line research
          assumed; i.e. one affected by stress, sonority, etc.

2) Today
   a) Briefly visit the “off-line” syllable via last Friday’s handout; but the gist is that off-
      line syllables are affected by multiple factors and seem to be probabilistic
   b) Coleman & Pierrehumbert (1997): build and test a probabilistic model of
      sylabification
   c) Zamuner et al (2004): phonotactic probability effects in children’s productions

Coleman & Pierrehumbert (1997) – Probable Cause

3) Purpose: obtain well-formedness judgments from humans and probability models
   from the lexicon and compare for best fit

4) Why Modeling?
   a) "We undertake to model productivity because it is a standard diagnostic for the
      psychological reality of abstractions. Modeling in detail the perceived well-
      formedness of neologisms provides us with an opportunity to assess how prosodic
      structure figures in the cognitive system." (p.2)

5) Human Task
   a) Stimuli:
      i) Nonsense items: blick vs. sphick vs. bnick
   b) Procedure:
      i) Participants judge how well-formed an item is on a scale of 0-12 (0 = “bad”)

6) What predicts those judgments?
   a) It can’t be orthodox phonology because judgments of non-words are gradient:
      sphick worse than blick, but better than bnick

Background on Probability

7) What’s the chance of getting a 6 on one die? 1/6
   a) p >= 0 and p <= 1
   b) Probability distribution requirement:
      i) the sum of the probabilities of all possible outcomes is 1
8) What's the chance of getting 6's on two dice? \( \frac{1}{6} \times \frac{1}{6} = \frac{1}{36} \)

9) How do we apply this to language? This type of unigram model certainly can't be enough:
   a) problem #1: \( p(bnIk) = p(b) \times p(n) \times p(l) \times p(k) \); if that were true, \( p(bnIlt) = p(bIlt) \)
   
   b) problem #2: \( p(mrup) < p(mrupation) \)

C&P’s Probability model(s)

10) C&P use a version of a probabilistic context-free grammar for syllables
   a) Each 'bit' of a nonsense item can be assigned a probability.

11) The model/grammar (sans irrelevant detail and straight to the gist):
   a) How it works:
      i) Take a bunch of one- and two-syllable words
      ii) Parse them into [O-R][O-R]
      iii) Calculate how frequent each bit is
      iv) Reserve a token amount for non-occurring items (smoothing; avoid zeros!)

   b) For example:
      i) \( p(bnIk) = p(bn) \times p(Ik) \)
      ii) \( p(Ik) = c(Ik)/c(\text{all rhymes}) \) - [small amount]
      iii) \( p(bn) = 1/c(\text{all non-occurring rhymes}) \) * [small amount]

   c) Results:
      i) \( p(\text{word}) \) YES
      ii) \( \ln(p(\text{word})) \) YES
      iii) \( p(\text{worst}) \) YES
      iv) \( p(\text{best}) \) NO!
12) R-code and English data for these on the website!

13) What does this mean?
   
   a) Note that the model instantiates some tricky aspects of syllabification that may not be right:
      i) E.g., the parser syllabifies all medial VCV sequences as V.CV. What to make of that?
      ii) The tricky thing about models is they can inform us only so far as how that particular model was instantiated
   
   b) What about phonology per se?
   
   c) What about the syllable?


14) Some previous claims:
   
   a) Coda production for children reflects probabilities too
   
   b) Codas are produced earlier in words with high phonotactic probability

15) Background
   
   a) Kids produce open syllables before they produce closed ones:
      i) E.g., 'man' [maen] -> [mae]
   
   b) CV syllables are less marked than CVC syllables
      i) YES: Languages with only CV
ii) YES: Languages with CV and CVC
iii) NO: Languages with only CVC

16) Methods
   a) Participants: children aged 1;8-2;4

   b) Materials:
      i) CVC nonwords and accompanying pictures of nonsense animals
         (1) based on CVC probabilities only overall or only wrt/ coda?
      ii) Neighborhood density controlled, but HOW?

   c) Procedure:
      i) Auditory presentation with picture: repeat item

17) Results:
   a) Children produce more codas correctly in high probability environments than in low ones, and also do the same for rhymes

   b) Effect of age: younger children produced quantitatively fewer responses overall but no qualitative differences

   c) No correlation with vocabulary size
      i) Other research finds such a correlation for older children; difference could be because children were so young that vocabulary size was relatively homogenous across the children

18) Implications:
   a) Young children’s representation of the input is much richer than classically believed.
      i) This correlates well with the findings of Z&O, i.e. the "orthographic effect" on pre-literate children.

19) Technical issues
   a) Bins for probabilities
   b) A number of No-Responses in the data but always noisy since these are kids!
   c) Non-parametric data, so aggregation for stats as in Treiman papers
   d) Lots of post hoc analyses to dismiss other analyses (fairly typical in the field)

20) Final word:
   a) "The results show that children do not just progress from deleting codas to eventually producing them (Jakobson, 1941/1968) or even that children do not just produce more frequent codas before infrequent codas (Stoel-Gammon,1998; Zamuner, 2003). Rather, the results illustrate that children differentially produce the same coda depending on the non-word’s phonotactic environment."

Hammond/Ohala-4