Problem Set 2: Finding Words (and other Stuff) with Statistical Learning
Due: When ready

In this assignment, we evaluate the effectiveness of **statistical learning via transitional probability** (SAN) as a mechanism for word segmentation. The idea of statistical learning via finding local minima of transitional probabilities was originally due to Chomsky (1955). But he noted "the problem is whether this [word segmentation via transitional probability] can be done on the basis of a corpus of a reasonable size' (section 45 footnote). It cannot, despite the overwhelming enthusiasm for SAN's findings.

You are given a text file already prepared for this experiment—by Tim Gambell, a Yale undergraduate student at the time. We did some work together which turned out to be the first critique of statistic learning: see **here** and **here**. That work, along with other critical voices, has led to a more sensible understanding of statistical learning.

The data was extracted from the mother's speech to children in Roger Brown's landmark study. Each utterance has already been transcribed into phonemes by the use of the CMU pronunciation dictionary. For instance, the very first sentence in the input is "big drum", which has been converted into:

```
bPih1PgPSWdPrPah1PmPSWU
```

Each phoneme is followed by **P**, each syllable is followed by **S**, each word is followed by **W**, and each utterance boundary is followed by **U**. Therefore, "m" in the transcription above is appended with PSWU as it is a phoneme, syllable, word boundary, and utterance boundary at the same time. The phonemes in the CMU pronunciation dictionary are transcribed using APARBET (https://en.wikipedia.org/wiki/ARPABET), and the numbers (0, 1, 2) mark stress with 1 for primary stress, 2 for secondary stress, and 0 for no stress.

Your implementation should take the input like above, remove the word boundaries for each utterance (as marked by U), and produce a segmentation based on statistical learning. The word boundaries your implementation inserts should be compared with the Ws in the original input. As usual, precision, recall, and F-score need to be reported.

Bonus Even further back, the proposal came from Harris (1953, *Language*), in a paper entitled "From phonemes to morphemes", but he proposed using transitional probability local minima over adjacent phonemes to discover morphemes. It turns out that the phoneme-to-morpheme idea works better—for morphemes. Play around a bit and extract the most frequently extreme phoneme sequences: Do they look like the morphemes of English? This idea was put to test, again by **a couple of undergraduate** students at the time: with some clever engineering, this simple idea beat many highly

complex morpheme segmentation systems at the time, and remains a benchmark that is very hard to beat for English.

Bonus Note that in order to gather transitional probabilities over syllables, the infant needs to syllabify the phoneme sequences into syllables. The principle to do so is MaxOnset, as many of you know, but in order to do that, you need to obtain a list of legal onsets. How would one do so? In this bonus problem, remove all the S's in the data and build syllables yourself before running the segmentation implement built earlier.

Bonus The computational modeling papers linked above were written quite some long time, and had some speculative nature. Later developmental work has essentially supported the general approach (see the Johnson and White 2019 review paper). The more comprehensive, and psychologically reasonable system was put forward by **Constantine Lignos (2012)** who, incidentally, was also a Yale student at an earlier point.