# The Model

We can expect to have a "language team" that includes computational and phonetic skills, but not specifically speech tech skills. Adapting the transcriber to the language is a one-time task performed by the language team, not by ultimate end user.

Epitran consists of preprocessing rules *Pre,* mapping rules $M$, and post-processing rules *Post.* We can expect the language team to write *Pre.* In the out-of-the box model, *Pre* is (probably) empty.

We provide a generic set of $M$, constructed from the Epitran tables. It pools all of the pronunciation mappings from all of the languages that use Latin scripts. Each mapping in $M$ takes a sequence of graphemes and maps it to a sequence of phones, representing the canonical pronunciation.

Let's put Post aside for now.

We also assume that the language team compiles a wordlist $V$. *Pre* can be applied in advance to obtain $V' = Pre(V)$.

Let $W$ be the word sequence. Out-of-vocabulary items can be covered by providing a trivial single-grapheme "spelling-mode" word for each grapheme.

A word sequence determines a unique grapheme sequence $G$. Applying the mapping converts $G$ to a "canonical pronunciation" phonetic sequence $P$. In general, the mapping from $G$ to $P$ is ambiguous, both because there may be multiple mappings that have the same phone on the lefthand side, and because there are competing mappings that segment $G$ in different ways.

Let $A$ be the perceived phones, that is, the output of the acoustic model. There is a string-edit transformation from $P$ to $A$, involving insertions, deletions, and substitutions. Specifically, a single phone in $P$ may be deleted, a single phone in $A$ may be inserted, or a single phone in $A$ may be substituted for a single phone in $P$.

*(Picture goes here)*

A "parse" is a sequence of words, each of which decomposes into a sequence of "beads" defined by the instances of mapping rules from $M$. The cost of a parse is the sum of the following: the word cost, the transition cost from bead to bead, and an emission cost which is the sum of the costs of deletions and substitutions that take the $P$ portion of the bead to its $A$ portion. Insertions are treated as separate, special beads that contain only an $A$ phone and no $G$ or $P$ elements.

In sum, the parameters are as follows:

| | |
|---|---|
| $I$ | insertion cost(s) |
| $D$ | deletion cost(s) |
| $S$ | substitution cost matrix |
| $T$ | transition costs |

The following are not parameters, but are necessary resources:

| | |
|---|---|
| $M$ | generic one provided, may be edited by language team |
| $Pre$ | provided by language team if desired |
| $V$ | vocabulary |

We fix $I$ and $D$ by fiat. As a first approximation, they are single parameters.

There are two ways of obtaining an estimate of $S$. We may use a phonetic similarity metric to define substitution probabilities, or we may use the $P$-to-$A$ table that Antonis has.

We use $V'$ to estimate the costs of transitions from grapheme to grapheme (that is, from one spelling pseudo-word to another). The transition probability from a true word to a grapheme is estimated by looking at word-initial graphemes in $V'$. The transition probability from grapheme to true word is estimated by looking at word-final graphemes in $V'$. The transition cost from true word to true word is defined to be zero.

To get the lowest-cost parse, given $A$ and possibly also $W$, we formulate the model as an HMM. With an HMM, there is a state at each position in the input sequence, and an output from each state. As a result, input and output sequence have exactly the same length.

In the way I have described a parse, the basic units are words and beads, and beads may cover multiple input elements and multiple output elements, and not necessarily the same number of each.

To shoehorn this into an HMM model, we introduce "continuation states" that represent partial beads and partial words. After seeing the first $A$ phone in a bead $B$, one option is a state $q_1$ that represents a partial match of $B$. The emission cost is the substitution cost for $P \rightarrow A$. The second $A$ phone may be reduced to state $q_2$, representing the second part of $B$. The transition cost for $(q_1, r)$ is infinite unless $r = q_2$, and the transition cost for $(r, q_2)$ is infinite unless $r = q_1$.