Disambiguation symbols are the symbols #1, #2, #3 and so on that are inserted at the end of phonemene sequences in the lexicon. When a phoneme sequence is a prefix of another phoneme sequence in the lexicon, or appears in more than one word, it needs to have one of these symbols added after it. These symbols are needed to ensure that the product L o G is determinizable. We also insert disambiguation symbols in two more places. We have a symbol #0 on the backoff arcs in the language model G; this ensures that G is determinizable after removing epsilons (since our determinization algorithm does remove epsilons). We also have a symbol #-1 in place of epsilons that appear on the left of context FST C, at the beginning of the utterance (before we start outputting symbols). This is necessary to fix a rather subtle problem that happens when we have words with an empty phonetic representation (e.g. the beginning and end of sentence symbols <s> and </s>).

We give the outline of how we would formally prove that the intermediate stages of graph compilation (e.g. LG, CLG, HCLG) are determinizable; this is important in ensuring that our recipe never fails. By determinizable, we mean determinizable after epsilon removal. **The general setup is: first, we stipulate that G must be determinizable. This is why we need the #0 symbols (G is actually deterministic, hence determinizable). Then we want L to be such that for any determinizable G, L o G is determinizable. [The same goes for C, with L o G on the right instead of G]. There are a lot of details of the theory still to be fleshed out, but I believe it is sufficient for L to have two properties:**

L^{-1}

- **must be functional**
  - **equivalently: any input-sequence on L must induce a unique output-sequence**
  - **equivalently: for any linear acceptor A, A o L is a linear transducer or empty.**
- **L has the twins property, i.e. there are no two states reachable with the same input-symbol sequence, that each have a cycle with the same input sequence but different weight or output sequence.**

The same applies of course to the C transducer. We believe that the transducers as our scripts and programs currently create them have these properties.