8a. HTK

The Hidden Markov Model Toolkit (HTK) is a speech recognition toolkit developed at Cambridge University. It is a set of programs that can be used to build speech recognition systems. Part of the process of building such systems involves force-aligning training data - i.e. automatically lining up phonemic-transcriptions of known words with the audio signal in the training recordings. LaBB-CAT takes advantage of this capability to facilitate forced-alignment for your transcripts.

In order to do this, HTK needs the following ingredients:

1. a set of recordings broken up into short utterances,
2. orthographic transcriptions of each utterance, and
3. phonemic transcriptions of each of the words in each utterance

You already have 1. and 2. - i.e. a set of recordings with transcripts that include the start and end times of each line.

You also mostly have 3. as well, if you have done a previous exercise which included generating a pronunciation layer generated using a lexicon like CELEX or the CMU Pronunciation Dictionary. However, there are words in your transcripts that aren't in CELEX, and so we will explore some mechanisms for filling in their pronunciations.

In this exercise you will

- install the HTK Layer Manager,
- provide some pronunciations that are missing,
- force-align the speech of one of the participants in your database, and
- check and manually correct the alignments.

> **!** Important
>
> In this exercise, you will set up Praat Integration in your web browser. There is currently no Praat integration support for Microsoft's Edge' browser, so if you normally use 'Edge' on Windows, you may need to swap to another browser for this exercise - e.g. *Google Chrome*, or *Mozilla Firefox*.

Installation

HTK is not free software in the "GNU" sense - i.e. we can not distribute it with LaBB-CAT, instead you have to download it yourself from the Cambridge University website – however it is free in the "no cost" sense, you just need to register on the HTK website, and you can then download and use HTK free of charge.

(1) Register at http://htk.eng.cam.ac.uk/register.shtml
(2) Download the version of HTK that is appropriate for the computer that LaBB-CAT is installed on:
    - For Windows systems, there are pre-compiled .exe files that you can download.
    - For Unix-like systems including OS X, you need to download the source code, which you will then install following the provided instructions.
(3) Unzip (for Windows) or compile and install (for Unix-like systems) the downloaded files on the computer that LaBB-CAT is installed on.

> **Note**
>
> If your LaBB-CAT server is installed in a *Docker Container*, LaBB-CAT can itself download and compile HTK (i.e. steps 2 and 3), as long has you have a username and password for the HTK website. In this case, these steps are automatically attempted when you install the HTK layer manager (below).

Now, you have to install the HTK Layer Manager, which is the LaBB-CAT module that provides HTK with all the data it needs, and then saves to alignments HTK produces back to your database.

(4) Select the *layer managers* menu option.
(5) Follow the *List of layer managers that are not yet installed* link at the bottom.
(6) Find HTK Manager in the list, and press its *Install* button, then press *Install* again. You will see a form with boxes that may already be filled in.
(7) The layer manager needs to know where the HTK programs have been saved, which is what you need to enter in the blank *Path to HTK tools* box.
    If this box is not blank, it means that LaBB-CAT has already found HTK for you, so you should leave the default value already set.

> **Note**
>
> The other two boxes you may see are for the username and password you registered on the HTK website. On some systems, if you have not already installed the HTK tools on your LaBB-CAT computer, providing the username and password allows LaBB-CAT to attempt to download and install HTK itself. This may not work in all cases, as LaBB-

CAT may not have the resources or privileges it requires to compile or install software.

(8) Press *Configure*.
You will see a window open with some information about the HTK Layer Manager.

(9) Now you need to add a *phrase layer* for the HTK configuration:

- **Layer ID:** `htk`
- **Type:** *Text*
- **Alignment:** *Intervals*
- **Manager:** *HTK Manager*
- **Generate:** *Never* (note that this may see counter-intuitive, but we will be running HTK manually rather than allowing it to run automatically when transcripts are uploaded.)
- **Description:** `HTK alignment time`

(10) When you configure the layer:

- Your *Pronunciation Layer* will be the *phonemes* layer you created in an earlier exercise.
- To the list of *Pause Markers* you should add a full-stop (period).
  This is because the exercise transcripts use . as a 'short-pause' marker, not an 'end of sentence' marker. i.e. your Pause Markers should be:
  – .
  (a hyphen, then a space, then a full-stop/period)
- The rest of the settings should be left with the default values.

Tip

If you're curious about what the configuration options do, hover your mouse over each option to see a 'tool tip' that describes what the option is for.

We configure the layer to 'never' generate, because we're going to trigger forced-alignments manually, one speaker at a time, once we're happy that the phonemic transcriptions are in place.

Missing Pronunciations

Now we're going to check the situation of our pronunciations on the phonemes layer more carefully…

(11) Go to the *transcripts* page and open the transcript *UC207YW*

(12) Tick the *phonemes* layer.
(13) Have a look through the transcript to see where the missing phonemic transcriptions are.

You'll see they divide into various broad types:

- Typos like "Febuary"
- Specialist or invented words like "tarseal"
- Contractions like "me's" and "thing's"
- Proper names like "Bealey"
- Possibly filled pauses like "um"
- Hesitations and interrupted words like "exac~", etc.

HTK needs a phonemic transcription for every word on a line in order to force-align that line. So every line where there's a gap on the phonemes layer would be ignored by the HTK layer manager.

(14) There's another problem in this transcript, which isn't necessarily immediately obvious.
Look for the hesitiation "w-" and the filled-pause "mm" to see if you can see what it is.

'False positives' from the lexicon will also play havoc with forced alignment, as HTK believes what it's told about the pronunciations given to it, and will do it's best to find an alignment that includes every phoneme.

Each of these problems needs to be addressed before we do forced alignment, although the solution for each will vary. Some involve improving the transcript, others will involve adding new words to our dictionary.

- For false-positives like "w-" and "mm", the easiest solution is to transcribe these differently. Hesitations like "w-" are discussed below. We will use "mmm" instead of "mm".
- For very short false-starts like "w-", the CELEX layer manager has been built to give a helping hand. In addition to looking up phonemic transcriptions in CELEX, it will also compute them for very short tagged false-starts. The tag it recognizes is a trailing tilde ~, so we need to change "w-" to "w~" etc. Then the CELEX layer manager will append a schwa to the initial consonant, and save that as the pronunciation (i.e. /ə/).
- For invented or misspoken words, or longer interrupted words, which we're not likely to see again in any other transcript like, "me's" and "exac~", we will add a 'pronounce' tag in the transcript, which includes the correct pronunciation. Again, the CELEX layer manager knows to check for pronounce annotations, and uses the given phonemic transcription instead of looking up the CELEX data.

- For proper names and contractions like “thing’s” that we’re likely to see over and over again in different transcripts, instead of tagging each one individually, we will add them to the dictionary of pronunciations that the lexicon layer manager looks up.

As you can see, the first three methods involve editing the transcript. This can be done by editing the original file in ELAN, and then re-uploading it into the database for processing.

Alternatively, LaBB-CAT has a mechanism for editing the transcript ‘in-situ’; this doesn’t update the original file, but it’s sometimes much more convenient, and this is the method we’ll use for this exercise.

(15) Click on the word “Febuary”, and select the *Edit Transcript* option from the menu.
A window will open that allows you to edit that line in the transcript.
(16) Correct the spelling of the word to be “February”
(17) Press *Update*
(18) Press *Close*
(19) Similarly change “w-” to be “w~” and “mm” to be “mmm”
(20) Click on the word “me’s”, and select the *Edit Transcript* option from the menu.
It seems unlikely that anyone else will say “me’s”, so instead of adding it to the lexicon, we’re simply going to tag this token with a *pronounce* tag. This is achieved by adding the pronunciation we want to tag it with in square brackets, using the DISC phoneme symbols. We have to add the *pronounce* tag immediately after the word, with no intervening space. (This transcription convention also works if you edit the original transcript in ELAN)
(21) Change the line text from “...bit of me’s a bit …” to be “…bit of me’s[miz] a bit …” instead.
(22) Click *Update* and then *Close*
The *pronounce* annotation you’ve just added isn’t displayed in the transcript. It’s added to the *pronounce* layer, which is for this type of manual pronunciation tagging.
(23) Scroll to the top of the transcript and tick the *pronounce* layer so that it will be displayed.
When the transcript is reloaded, you will be able to see “miz” as an annotation on “me’s”, and that this has been copied into the *phonemes* layer by its layer manager.
(24) Similarly you should tag the word “exac~” with the pronunciation `Igz{k`

This method takes care of instances where the transcript is incorrect, and ‘one off’ missing pronunciations. However, for missing words that are likely to appear over and again in the corpus, including names like “Bealey”, “Wainoni”, and “Lyttleton”, and filled pauses like “mmm”, “um”, etc. it’s not efficient to tag every token. Instead, we add these to the lexicon.

(25) Select the *participants* menu option.
(26) Find *UC207YW* in the list, and click on their name.
(27) Click the *All Utterances* link below their attributes.

(28) Leave the default selections and press *List*.
This displays a page with a list of all the speaker's utterances, from which you can do various things with all the utterances of a particular participant in the database.

(29) Press the *Htk* button.
You will see a progress bar while LaBB-CAT identifies missing words. Then a page will appear that lists unknown words.

(30) Basically you need to fill in the boxes with the pronunciations and click *Save Pronunciations*.

Note

- You don't have to fill them all in at once, you can do a few, and click *Save*, which will save your work and list what's left.
- You don't have to fill them all in, you can leave some empty and continue with the HTK forced-alignment by clicking *Start* (HTK will ignore any lines where the remaining unknown words appear, but the ones you filled in will be included).
- Some of the boxes will be initially filled in with a suggestion from the lexicon layer manager - these may or may not be correct, and aren't saved until you save them.
- The pronunciations have to be in the 'DISC' format - i.e. one character per phoneme, with no spaces. There's a 'helper' link on the right of each pronunciation box - if you click it, it expands into a list of clickable phonemes - just the ones that aren't ordinary letters, and diphthongs etc.
- The *search* button lets you look up the lexicon for similar words - this probably won't help for place names, but for words like "tarseal", you can click the *lookup* button, enter "tar seal" in the box as two separate words, and you'll get back the DISC pronunciation of each word, with clickable buttons to copy the given pronunciation into the box. This is useful for digits and numbers too, which may not be in the lexicon - so for "1", search for "one" and copy the pronunciation.
- If you click on the word itself, the transcript for the first instance of that word is opened, in case you want to listen to it, or in case it's actually just a typo and you want to correct the transcript.
- If you're using CELEX, when you specify the pronunciations, it's recommended to put syllable separators (-) and primary stress markers (') too - e.g. for "tarseal" you can put *t#sil* but it would actually be better to put *t#-'sil*. These markers are entered into the dictionary even though they're stripped out for HTK, and they may come in handy later (e.g. the syllable separators are used by the CELEX layer manager to count syllables).

When you add pronunciations this way, they're added to the dictionary and all the instances of those words in LaBB-CAT are updated with the pronunciations - not just the participant you're looking at, but all participants in the database. So you only have to come up with a pronunciation for each word once.

Alignment

(31) Once you've filled in all the missing pronunciations, forced alignment will start automatically. If you want to start forced alignment before you've entered all pronunciations, click the *Start* button at the bottom of the page.

You should see a progress bar while the forced alignment is running. It will take a few minutes to complete.

Once HTK has produced the word and segment alignments, it:

- sets the start/end times of the words on the transcript layer accordingly,
- adds new phone annotations to the *segments* layer with the alignments of the phones, and
- saves a timestamp in the *htk* layer.

When the layer manager has finished, you'll see a message saying "Complete - words and phones from selected utterances are now aligned."

Inspection/Correction

You can inspect/correct alignments using LaBB-CAT's integration with Praat.

(32) Go back to the *UC207YW.eaf* transcript.
(33) Tick both the *htk* layer and the *segments* layer.
You will see which lines have been force-aligned, as they have an HTK timestamp, and have the *segments* layer filled in. If it has missed some lines, this is most likely because there is an unknown word, another speaker speaking at the same time, or possibly HTK simply failed to align the line (there are various reasons this happens, including not enough data for training, noisy recordings, inaccurate transcription, etc.).

The interactive transcript page doesn't show you the alignments of the words or phones, but you can see those using Praat. You can open individual utterances in Praat directly from the transcript page, but first, the LaBB-CAT/Praat integration has to be set up; this only has to be done once:

i. On the top-right of the page, above the playback controls, there's a Praat icon - click it.
ii. Follow the instructions that appear (these vary depending on what web browser you use).
You may be asked whether to allow the "LaBB-CAT Integration Applet" to run. If you tick the "Do not show this again" option, then this message will not appear every time you open a transcript.
You may need to grant a browser extension permission to install, and it's possible you will need a connection to the internet in order to download this extension.
You also may be asked where Praat is installed; Navigate to the location where Praat is installed, and double-click the "Praat.exe" file (on some systems the file may simply be called "Praat"). The Praat program may open, and then immediately close, as LaBB-CAT tests it can communicate with Praat.

Now Praat integration has been set up, and you should be able to access Praat options in the transcript page from now on…

(34) Click on a line that has been aligned, and select the *Open Text Grid in Praat* option on the menu.
You may be asked you if want to allow access to the "LaBB-CAT Integration Applet" - if so, tick "Do not show this again", and click *Allow*.

Praat should open, and show you a spectrogram of the line's audio, with a TextGrid below that includes the words and the segments.

(35) If you click on a word, and hit the `tab` key, the word's interval is played. Try out various words, and see what you think about how accurate HTK has been with its alignment. Try this out with different lines in the transcript.
You will see that in some cases the alignment is pretty good, and in other cases, it's not so good. In the not-so-good cases, see if you can figure out why HTK got it wrong.

You may have noticed that, each time you open an utterance in Praat, a button appears in the transcript to the left of the line, labelled *Import Changes*. This button allows you to save any adjustments you might want to make to the alignments back into the LaBB-CAT database.

(36) If you feel confident using Praat, open an utterance TextGrid, adjust the alignments of the words an phones so that they're more accurate, and then click the *Import Changes* button in the transcript.

> ⚠ Warning
>
> These changes are flagged as manual edits, so if forced-alignment is run again, they will not be over-written with new bad alignments. Therefore it's important that the changes you make are actually improvements, because HTK will never change them again.

> ❗ Important
>
> There are some rules about what you can change:
>
> - You're not allowed to add or delete words (if this is necessary, it should be done by correcting the transcript instead).
> - All the phones must be within the bounds of their own word.
> - The start of the first phone should line up with the start of the word, and the end of the last phone should line up with the end of the word.
> - You should not change the alignment of the utterance itself (which would only be possible if you select the *Open Text Grid incl. ± 1 utterance in Praat* option).

In this exercise, you have seen how HTK can be used to compute word and phone alignments automatically from your data, but that there is a fair amount of careful transcription, tagging, and dictionary filling required. Even after all that work, perfect automatic alignments are not guaranteed, but LaBB-CAT has a mechanism for manually correcting poor alignments.

All this manual annotation and correction means a lot of work, but obviously somewhat less work than would be involved in aligning by hand the transcripts from scratch!