# FORCED ALIGNMENT OF DIFFERENT LANGUAGE VARIETIES USING LABB-CAT

Robert Fromont

University of Canterbury
robert.fromont@canterbury.ac.nz

## ABSTRACT

Large-scale phonetic study using speech corpora generally requires automated forced alignment of phones. When pre-trained acoustic models and pronunciation dictionaries are available for the language variety being studied, this is fairly straightforward. For other varieties, a number of other approaches can be used, including training acoustic models on the corpus data itself, adapting existing dictionaries, creating new ones, or inferring phonology from orthography.

This paper describes a number of approaches that can be used to successfully force-align different varieties of English, and other languages for which models and pronunciation dictionaries are not easily available, using LaBB-CAT.

**Keywords:** Forced alignment, pronunciation dictionary, pronunciation derivation, language varieties

## 1. INTRODUCTION

Phonetics research using large speech corpora generally requires automated 'forced alignment' to identify phones and their start and end times. There is a growing number of automatic speech recognition (ASR) tools that can help with this process, which generally use acoustic models, a pronunciation dictionary, and orthographic transcriptions of the speech to arrive at phone alignments.

Several ASR tools are readily available, such as the HMM Tool Kit (HTK) [27], the Munich Automatic Segmentation system (MAUS) [21], and Kaldi [19]. Around these are often wrapped higher-level systems that either include pre-trained models, like the Penn Phonetics Lab Forced Aligner (P2FA) [28] and WebMAUS [24], or provide functionality that eases data preparation and processing, like the Montreal Forced Aligner [17] and LaBB-CAT (described briefly in section 2 below, and in detail in Fromont & Hay [10]).

Pre-trained acoustic models work well aligning data that is similar to the speech on which the models were trained, but their effectiveness is decreased when the the data to align is not similar to the training data; for example models trained on American English speech will perform well when aligning American English, but may be less effective on other varieties of English [11].

When pre-trained acoustic models are unavailable there are a couple of possible approaches:

- pre-trained models for some other language or variety may still be used, possibly followed by some manual correction, or
- if there's enough speech data[1], acoustic models can be trained directly on the data to be aligned.

Similarly, the pronunciation dictionary used can influence the quality of alignments, as phonemes that are present in the speech may be absent in the dictionary or vice versa; for example a rhotic English dictionary used to align a non-rhotic variety of English will result in spurious /ɹ/ phone alignments.

For major language varieties, pronunciation dictionaries are readily available, but they represent 'standard' pronunciations, and do not describe regional accent differences within their respective nations, nor varieties spoken in other countries. For such varieties, there are a number of options available for arriving at a pronunciation dictionary:

- the dictionary of a similar variety can be used or adapted,
- it may be possible to automatically generate a dictionary for the specific variety,
- a dictionary can be created from scratch manually, or
- in some cases, pronunciations can be inferred directly from the words' orthographic spelling.

Each of these approaches to forced alignment can be applied using LaBB-CAT. In the following section I will very briefly describe LaBB-CAT's general functionality, and then in the rest of the paper I will describe how it can be used to apply each of these approaches to forced alignment.

## 2. LABB-CAT

LaBB-CAT is a browser-based corpus management system, developed at the New Zealand Institute of Language, Brain and Behaviour at the University

of Canterbury, originally designed for sociophonetic research in the Origins of New Zealand English (ONZE) project [9]. It has since been used to manage a number of different corpora at different institutions (for examples, see [4, 13, 14, 15, 18, 23, 25]).

It is designed to be a data store for speech recordings, orthographic transcripts, and linguistic annotations. Annotations are organised in 'layers', and can be manually applied or automatically generated. Annotations can include participant or transcript meta-data, broad temporal partitions of the recordings, word-level tags, and sub-word units such as time-aligned syllables and phones.

Annotations can be searched for patterns, and exported in a variety of formats for further analysis.

Automatic generation of annotations is performed by 'layer managers', which are modules designed to take inputs such as recordings, other annotation layers, and dictionaries, and perform computations that produce more annotations (and possibly other kinds of data such as frequency lists, etc.).

Layer managers exist to perform a wide variety of annotation tasks, including tagging tokens matching patterns, summarising chunks with count and rate information, tagging and analyzing word frequencies, syntactic parsing, tagging tokens with lexical information, and forced alignment.

The user interface facilitates import, visualisation, and export of data, dictionary management, bulk acoustic measurement with Praat [3], and general corpus management.

LaBB-CAT is free, open-source software.

### 2.1. HTK Layer Manager

The layer manager commonly use for forced-alignment in LaBB-CAT is the 'HTK layer manager', which integrates LaBB-CAT with HTK [27]. Although this layer manager is included in LaBB-CAT, the HTK software it integrates with must be downloaded separately[2].

Usually this module uses a 'train-and-align' approach to forced alignment, using four phases:

1. A phonemic transcription annotation layer tags each word token with its pronunciation. This is done by some other layer manager (e.g. the 'CELEX English layer manager', described below).
2. Utterances and orthographic transcripts are extracted, and a pronunciation dictionary is compiled from the phonemic transcription annotation layer. This is generally done per speaker.
3. Speaker-dependent monophone models are trained using the data gathered in step 2.
4. The data is force aligned using the newly-trained models, and the resulting aligned phones are saved to the 'segments' layer.

Other possible configurations are described in the following sections.

## 3. ENGLISH

### 3.1. Major Varieties

For major varieties of English, the most straightforward option is to use a standard dictionary for generating the phonemic transcription layer (step 1 in the HTK layer manager process above).

#### *3.1.1. British English*

For 'British English' speech data, the CELEX English [1] lexicon can be purchased and downloaded from the LDC[3] and LaBB-CAT includes a module designed specifically to integrate with it; the 'CELEX English layer manager'. Once the CELEX files have been loaded into the layer manager, it can be configured to tag word tokens with any lexical information in CELEX, including phonemic transcriptions[4].

#### *3.1.2. American English*

For 'American English' speech data, the Carnegie Mellon University Pronouncing Dictionary (CMU Dictionary) [20] is freely available and again LaBB-CAT includes a module for its integration; the 'CMU Pronouncing Dictionary layer manager'. As there are no restrictions on distribution of the CMU Dictionary file, the layer manager includes the dictionary pre-installed[5].

In the case of American English and the CMU Dictionary, an alternative configuration of the HTK layer manager is possible: the acoustic models that form part of P2FA [28] are included in the HTK layer manager[6], which can be configured to use them instead of training new models (step 3 in the HTK layer manager process above). As these models were trained on American English, the resulting alignments are generally of high quality.

### 3.2. Other Varieties

The following subsections describe ways of dealing with other varieties of English.

*3.2.1. Using a dictionary of a similar variety*

Phonemic tagging and forced alignment were two functions implemented early in LaBB-CAT's development, when its primary use was for the ONZE project[7].

It was found that that the phonology of the 'British English' represented by the CELEX English lexicon is sufficiently similar to New Zealand English (NZE) that the resulting phone alignments are of acceptable accuracy. The phonemic labels are sometimes not ideal, but the differences are systematically identifiable (for example words like "systematically" have /ɪ/ as the final phoneme in CELEX, where in NZE ending with /iː/ would be more accurate), so it's not difficult to take them into account when identifying tokens at scale.

One problem is that the ONZE corpus contains a large number of New Zealand specific lexical items that are not present in the CELEX lexicon. This is a problem for forced alignment, because all words must have an entry in the pronunciation dictionary. However before forced alignment begins, LaBB-CAT identifies all words with no phonemic transcription, and presents a list, allowing the missing pronunciations to be filled in directly. The given pronunciations are added to LaBB-CAT's internal lexicon so they can be used for other speakers, avoiding duplication of work.

In this way, the ONZE Project has built up a supplementary lexicon of almost 17,000 entries. This includes a great number of proper nouns (e.g. "Christchurch") and Māori loanwords (e.g. "kōrero"), but also possessive forms (e.g. "Christchurch's"), words spelled with digits (e.g. "1972"), and words that have come into common use since CELEX was compiled (e.g. "blog").

So for NZE, using a supplemented CELEX lexicon has worked well for forced alignment. This approach has also been applied successfully with other varieties of English, including West Australian English [5], Liverpool English [25] and Glaswegian English [23].

*3.2.2. Generating a variety-specific dictionary*

The Unisyn lexicon [7] developed by Fitt can provide more variety-specific pronunciations; it includes an 'accent independent' lexicon for English and a set of scripts that apply rules to produce accent-specific lexicons. Rule sets are included for a number of British regional varieties (RP, Leeds, Edinburgh, Aberdeen, Cardiff, Abercrave, and County Clare), three American accents (General American, South Carolina, and New York), and two other countries (Australia and New Zealand). The accent rules can also be manually adapted to suit other varieties of English [8].

LaBB-CAT includes the 'Unisyn layer manager', which is designed for ingesting Unisyn accent-specific lexicons. Unisyn must be downloaded separately, and the included scripts executed to produce a lexicon for the desired variety. The resulting file can be added to LaBB-CAT, and then the layer manager can be configured to use it for the phonemic transcription annotation layer generated in step 1 in the HTK layer manager process.

This approach was used with the Edinburgh accent-specific lexicon by Solanki [22], and also for Hawai'i English data, using a 'General American' dictionary, by Drager et al. [6].

## 4. OTHER LANGUAGES

There are also a number of approaches that can be used to force align languages other than English.

### 4.1. CELEX

CELEX includes not only an English lexicon, but also a lexicon for German and another for Dutch. LaBB-CAT includes layer managers designed to integrate with those lexicons – the 'CELEX German layer manager' and the 'CELEX Dutch layer manager' – which work similarly to the English one in allowing words to be tagged with lexical information, including phonemic transcriptions which can then be used for forced alignment.

### 4.2. Custom Lexicons

For languages with no pronunciation dictionary available, it's possible for the researcher to compile their own dictionary; LaBB-CAT includes the 'Flat File Dictionary layer manager', which is designed to allow the upload of one or more text files in comma-separated-value (CSV) format which is then used as a lexicon for tagging word tokens. Such files can be compiled using commonly-available spreadsheet software, with one or more lines per word, and including a column for the word orthography and further columns for other lexical information, including pronunciation in any desired encoding (DISC, ARPABet, Unicode IPA, X-SAMPA [26], etc.).

Such a custom lexicon needn't be an exhaustive list of words in the target language, it only needs to include all the words used in the corpus, which may be relatively few in the case of read-speech corpora. LaBB-CAT facilitates the task by including a mechanism for automatically identifying words with

no phonemic transcription, and exporting a CSV file which can then be manually filled in and re-uploaded to the Flat File Dictionary layer manager.

Once the dictionary is complete and words have been tagged with their phonemic transcriptions (step 1 in the HTK layer manager process), forced alignment can continue as normal.

This approach has been successfully used by Kaźmierski [14] in compiling an IPA-encoded lexicon for Polish, followed by forced alignment using the train-and-align approach. It was also used by Heyne [12] to compile an ARPABet-encoded lexicon for Tongan. This was read speech, and there was not enough speech data for each participant to successfully use the train-and-align method, so Heyne used the pre-trained P2FA models (trained on American English) to achieve approximately correct alignments, which were then hand corrected using LaBB-CAT's integration with Praat.

The Flat File Dictionary layer manager can also be used in cases where a dictionary file for the language is available, but LaBB-CAT doesn't include another layer manager specifically designed for it.

### 4.3. Inferred Phonemic Transcription

The orthography of some languages is closely related to its phonology. For example, in the case of Te Reo Māori, it's possible to devise a relatively simple mapping from spelling to phonemes; most letters can represent themselves, with some letter clusters mapping to specific phonemes (e.g. "ng" → /ŋ/, "wh" → /f/, etc.). LaBB-CAT's 'Character Mapper layer manager' can be used to define the details of such a mapping, in order to generate the phonemic transcription annotation layer (the HTK layer manager's step 1). Once that's done, the rest of the forced alignment process can proceed as normal.

This approach has been successfully used to force align the MAONZE corpus [15], which is a bilingual corpus including both Te Reo Māori and New Zealand English. As LaBB-CAT allows configurable metadata to be applied to transcripts, each transcript was tagged with its primary language; "mi" for Te Reo Māori and "en" for NZE. For phonemic transcription tagging, the Character Mapper layer manager was configured to tag only "mi" transcripts, and the CELEX English layer manager was configured to tag only "en" transcripts.

Many MAONZE recordings in fact contain a mixture of the two languages, which were teased apart for phonemic transcription purposes by using LaBB-CAT's 'language' annotation layer (which allows tagging of multi-word stretches within the transcript). Where the primary language for the recording was marked as "mi", the individual phrases that were in English were annotated on the 'language' layer with the label "en", and primarily English transcripts featuring Māori phrases were similarly tagged. In this way, the two layer managers can correctly tag all and only their respective tokens, even in cases where the languages are mixed together in the same transcript.

Another language for which a set of rules can be easily map orthography to phonology is Spanish. LaBB-CAT includes the 'Spanish phonological transcriber layer manager' which uses rules implemented by Baytukalov [2] to tag Spanish transcripts.

### 4.4. BAS web services

WebMAUS [24] is a web service that force aligns uploaded speech recordings by using models trained on a wide variety of languages and accents, including American, Australian, British, and New Zealand English, and Hungarian, Japanese, Maltese, among other languages. WebMAUS has been made available as one of the BAS CLARIN web services [16], which also include a service called "G2P" which transforms orthographic transcripts into phonemic transcriptions using more sophisticated algorithmic methods than the rule-based ones mentioned in the previous subsection.

LaBB-CAT includes a module designed to communicate with these two web services, called the 'BAS Web Services Manager'. If the speech data is in one of the many languages supported by the services, and there are no restrictions preventing uploading the data to a third-party service, it's possible to phonemically transcribe and force align corpora managed by LaBB-CAT without using the HTK layer manager, and the other layer managers mentioned above, at all. This widens the number of languages that can easily be force aligned to include several for which pronunciation dictionaries and acoustic models are not otherwise available, and also removes the need to have 'enough data' for training models from scratch.

## 5. CONCLUSION

Forced alignment of large corpora of both major language varieties, and of others, is becoming more practical as the number of available tools increases.

LaBB-CAT is one such tool, which can marshall a number of the others tools, and provides functionality to streamline common tasks, such as dictionary filling and acoustic model training, to decrease the drudgery sometimes associated with this task, before phonetic investigations at scale can be done.

## 6. REFERENCES


[1] Baayen, H., Piepenbrock, R., Rijn, H. V. 1995. *CELEX2*. Linguistic Data Consortium, University of Pennsylvania Philadelphia.

[2] Baytukalov, T. 2015. spanish-pronunciation-rules-php. https://github.com/easypronunciation/spanish-pronunciation-rules-php.

[3] Boersma, P., Weenink, D. 2016. Praat: doing phonetics by computer [computer program]. http://praat.org/.

[4] Clark, L., MacGougan, H., Hay, J., Walsh, L. 2016. "Kia ora. This is my earthquake story". Multiple applications of a sociolinguistic corpus. *Ampersand* 3, 13–20.

[5] Docherty, G., Gonzalez, S., Mitchell, N. 2015. Static vs dynamic perspectives on the realization of vowel nuclei in West Australian English. *ICPhS*.

[6] Drager, K., Kirtley, M. J., Grama, J., Simpson, S. 2013. Language variation and change in hawai'i english: KIT, DRESS, and TRAP. *University of Pennsylvania Working Papers in Linguistics* 19(2), 6.

[7] Fitt, S. 2000. Unisyn Lexicon Release, version 1.3. http://www.cstr.ed.ac.uk/projects/unisyn/.

[8] Fitt, S. 2001. *Documentation and User Guide to Unisyn Lexicon and Post-Lexical Rules*. Centre for Speech Technology Research, University of Edinburgh.

[9] Fromont, R., Hay, J. 2008. ONZE Miner: the development of a browser-based research tool. *Corpora* 3(2), 173–193.

[10] Fromont, R., Hay, J. November 2012. LaBB-CAT: an Annotation Store. *Proceedings of Australasian Language Technology Association Workshop*. Australasian Language Technology Association 113–117.

[11] Fromont, R., Watson, K. 2016. Factors influencing automatic segmental alignment of sociophonetic corpora. *Corpora* 11(3), 401–431.

[12] Heyne, M. 2016. *The influence of first language on playing brass instruments : an ultrasound study of Tongan and New Zealand trombonists.* PhD thesis University of Canterbury.

[13] Jannedy, S. 2010. The Usage and Distribution of "so" in Spontaneous Berlin Kiezdeutsch. *ZAS Papers in Linguistics (ZASPiL)* 52(2).

[14] Kaźmierski, K., Kul, M., Zydorowicz, P. In press. Educated Poznań speech 30 years later. *Studia Linguistica Universitatis Iagellonicae Cracoviensis*.

[15] King, J., Maclagan, M., Harlow, R., Keegan, P., Watson, C. 2010. The MAONZE Corpus: Establishing a Corpus of Maori Speech. *New Zealand Studies in Applied Linguistics* 16(2), 1–16.

[16] Kislera, T., Reichelb, U., Schiela, F. September 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45, 326–347.

[17] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M. 08 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. 498–502.

[18] Pope, C., Davis, B. H. 2011. Finding a balance: The Carolinas Conversation Collection. *Corpus Linguistics and Linguistic Theory* 7, 143–161.

[19] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K. 2011. The Kaldi Speech Recognition Toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

[20] Rudnicky, A. 2014. Carnegie Mellon University Pronouncing Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[21] Schiel, F. 2004. MAUS Goes Iterative. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)* Lisbon, Portugal. 1015–1018.

[22] Solanki, V. J. 2017. *Brains in dialogue: investigating accommodation in live conversational speech for both speech and EEG data.* PhD thesis University of Glasgow.

[23] Stuart-Smith, J., Jose, B., Rathcke, T., Macdonald, R., Lawson, E. 2017. *Changing sounds in a changing city: An acoustic phonetic investigation of real-time change across a century of Glaswegian*.

[24] T. Kisler, F. S., Sloetjes., H. 2012. Signal processing via web services: The use case WebMAUS. *Proceedings of Digital Humanities 2012* Hamburg, Germany. 30–34.

[25] Watson, K., Clark, L. 2017. *The Origins of Liverpool English*.

[26] Wells, J. 1995. *Computer-coding the IPA: a proposed extension of SAMPA*. University College, London.

[27] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchovaltchev, , Woodland, P. 2006. *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department.

[28] Yuan, J., Liberman, M. July 2008. Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics '08* Paris, France. 5687–5690.


---

[1] Fromont and Watson [11] found that 'enough data' is approximately 5 minutes of speech, for English data

[2] http://htk.eng.cam.ac.uk/register.shtml

[3] https://catalog.ldc.upenn.edu/LDC96L14

[4] CELEX's 'DISC' phoneme set is used, described in section 2.4.1 of [1].

[5] The CMU Dictionary uses the ARPAbet phoneme set: http://www.speech.cs.cmu.edu/cgi-bin/cmudict\#phones

[6] The inclusion of the P2FA acoustic models in LaBB-CAT is by kind permission of Jiahong Yuan

[7] The LaBB-CAT software was originally called "ONZE Miner" because of this original close association with the ONZE project.