SCOLAR Tools Documentation

by Jonathan Yip

August 7, 2020

This manual provides details about the methods used in the SCOLAR Project to process audio recordings and their corresponding annotation files. In the project, orthographic transcriptions of audio data were submitted to a sequence of Praat and R codes, as well as the auto-segmentation program SPPAS (Bigi, 2018), to produce sets of word- and syllable-based phonetic transcriptions of Hong Kong Cantonese language productions. This documentation describes how to format annotation files for auto-segmentation, how to carry out phonetic forced alignment, and how to extract data files that can be further annotated for use in the CLAN program within the CHILDES child language database.

Contents

1. Getting Started

1.1 Software installation

The software required to run all of the code for processing are as follows:

- Praat (Boersma & Weenink), version 6.1.16 or later: https://www.fon.hum.uva.nl/praat/
- R statistical computing software, version 4.0.2 or later: https://www.r-project.org/
- SPPAS automatic annotation and speech analysis software (Bigi, 2018), version 1.9.8: http://www.sppas.org/

Praat can be installed anywhere on your system, so long as it has the correct file associations in your computer's program registry. On a Windows system, it's helpful to place this program in the Program Files folder on your C: drive so that any user on your system can access the application.

R comes with an installation wizard that helps the user install the program correctly onto his/her computer. For ease of use, an auxiliary program, such as RStudio (https://rstudio.com/products/rstudio/download/), can be used to modify and run R code more efficiently. All R code files carry the file extension *r. Use of R code in this project requires pre-installation of a set of R packages. The packages required by the R code are *readr* and *stringr*. These packages are downloadable using R's **Install package(s)…** option under the **Packages** menu. The most common issue with installing packages in R is that dependent packages have not yet been installed. To fix this, you should install any other packages that R reports are missing.

SPPAS also comes with an installation wizard and is available for Windows, MacOX, and Linux. However, the software requires both administrator rights on your system as well as the installation of Python 3.6 or later. Instructions on how to do this are shown at the following URL: http://www.sppas.org/installation.html.

1.2 File preparation

The files to be processed need to be in the correct formats. Audio recording files (in **.wav* format) require corresponding annotation files (in **.TextGrid* format) in order to be processed correctly. Before processing, the **TextGrid* files need to contain annotation intervals that correspond to each utterance produced in a recording. Different speakers in the recording should have intervals placed onto separate interval tiers. Note that the names of these tiers is unimportant, but the order matters: mother talkers are placed on the first tier, and additional talkers, such as child talkers, are placed below that tier. Each tier should contain comprehensive set of intervals that have the correct labels for the orthographic representation of the utterance, as well as correct onset and offset boundary times. In a given interval, there should be as little extraneous audio material (noise or silence) as possible, except for any silence that is relevant to the signal, such as a brief interval of silence during an oral-stop constriction, or longer durations of silence duration a pause between words or phrases.

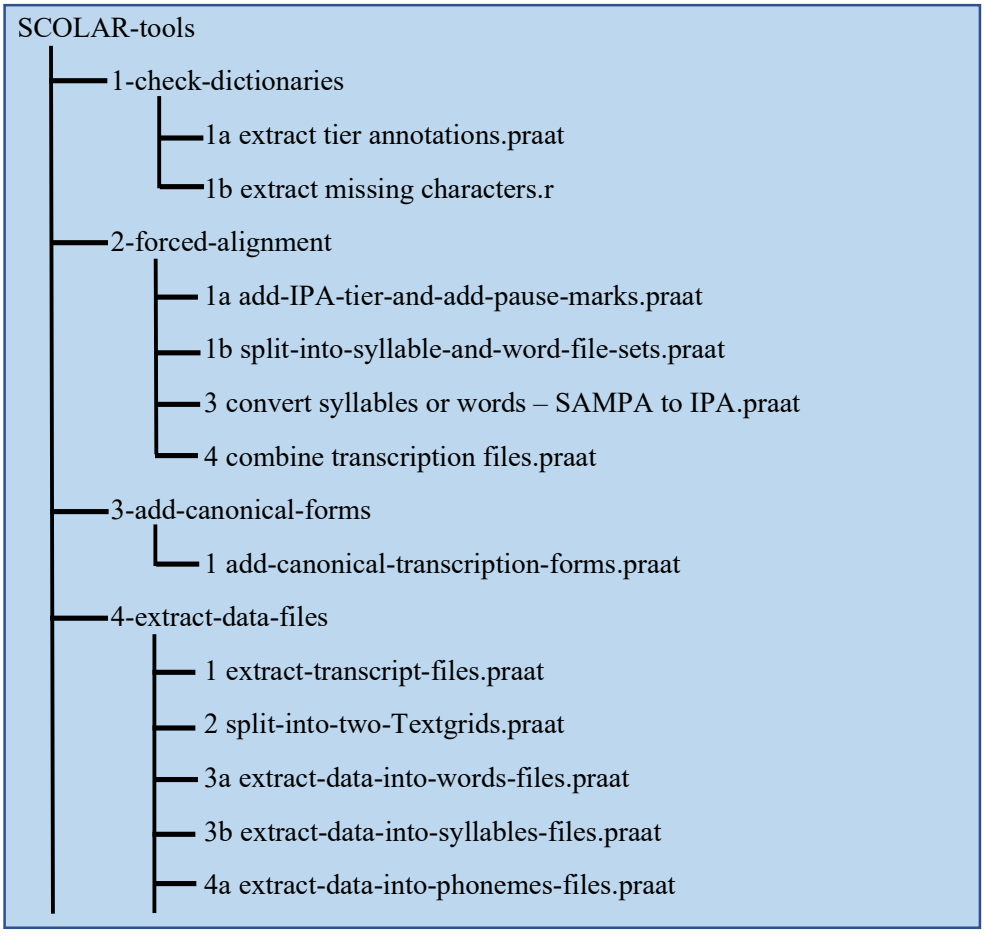
Within each annotation, labels may contain common punctuation symbols such as commas (,), full stops/periods (.), and question marks (?). However, SPPAS does not tolerate use of punctuation marks formatted for Chinese script such as ，, 。, and ？. So you should check all instances of punction and make sure the symbols have the appropriate format. The comma has a special use in the code in that it is used to indicate a speech pause. The plus sign (+) is also an indicator of such pauses, and so the comma and plus sign can be used interchangeably in the transcriptions. Pauses (commas or plus signs) should only be indicated if a speech pause is truly contain within a given interval, since indication of a pause that isn't actually there will result in a pause being forced aligned with continuous speech without pauses.

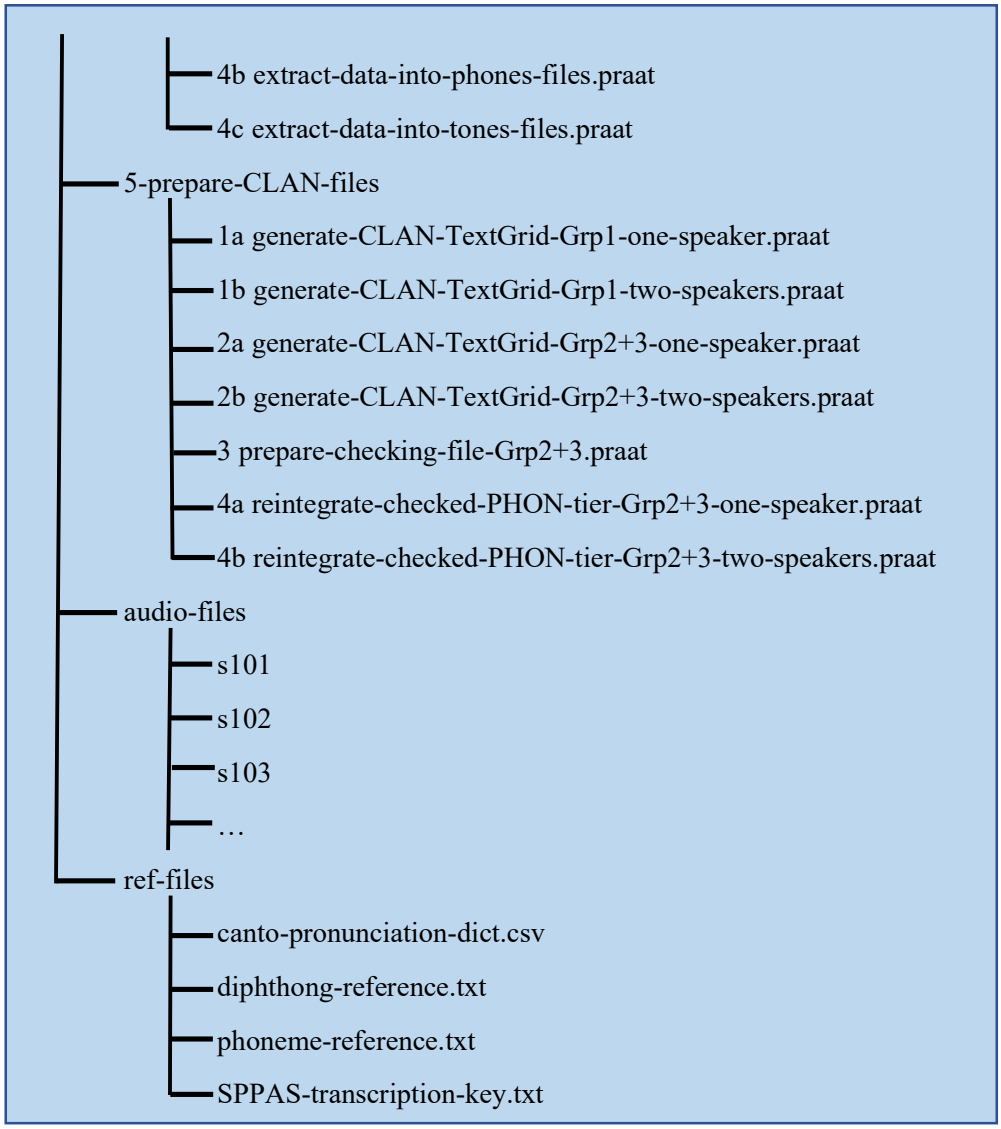
Additionally, non-linguistic sounds in the signal, such as laughing or coughing, cannot be contained in the annotation intervals, so attention should be paid in making sure that annotation intervals omit these types of non-linguistic sound productions.

It is important to note that only one voice per recording can be processed at a time. Because of this, it's important to split files into annotations for the first talker (*-1.TextGrid) and annotations for the second talker (*-2.Textgrid). For recordings with single-talker annotations, both the audio file and annotation file can be copied and then renamed with an appended *-1* suffix, e.g. *s101alf-1.wav* and *s101alf-1.TextGrid* for the ADS film-task recording for subject s101. For recordings that contain annotations for two talkers, new files should be named accordingly. Since only the mothers' voices has been processed, we have only copied the audio file for the first talker ("Mother"). However, a new TextGrid file for only the first talker should be made (using the **Extract one tier…** option in Praat) and named with the *-1* suffix, as well as a TextGrid file for only the second talker (child), named with the *-2* suffix. Original files can be kept in the subdirectory in the folder *audio-files* corresponding to the subject, and files renamed for the first talker (Mother) can be placed into the folder for Step 1 (*1-check-dictionaries*) for processing (refer to Sections 1.3 and 2.1). Child-talker annotation files with the *-2.TextGrid* suffix should be placed in the same talker subdirectory under *audio-files*.

1.3 File locations

For convenience, a template file structure containing all necessary code has been provided in the *SCOLAR-tools* software package. The structure of the package is as follows:





Shortcuts to dictionary (*.*dict*) and vocabulary (*.*vocab*) files under 1-check-dictionaries need updating so that the path goes to the appropriate directory on your system. Dictionary file shortcuts should point to the *resources* > *dict* subdirectory of your SPPAS installation, and vocabulary files are found in the *resources* > *vocab* directory of SPPAS. A shortcut to the SPPAS run file (*sppas.bat*), found under *2-forced-alignment*, should be linked to the **.bat* file of the same name found in the main installation directory for SPPAS. The shortcut for *canto-pronunciation-dict.csv* should be linked to the location of this file in this package's *ref-files* folder.

Additionally, you will need to add the latest version of our dictionary files to SPPAS's set of resource files. The files *yue-monosyllabic.dict* and *yue-polysyllabic.dict* are provided in the base directory of the package and should be moved to the *SPPAS* > *resources* > *dict* folder. Updated vocabulary files *yue.vocab* and *yue_chars.vocab* files are also provided and should be placed in the folder *SPPAS* > *resources* > *vocab*. All four of these files will require constant updating when new phonetic forms are encountered in the corpus during initial transcription.

2. Checking Dictionary Resources

2.1 Extracting tier annotations

The initial step in ensuring that annotation files can be forced aligned is to check the files for words that occur in the provided dictionaries. The code in the directory *1-check-dictionaries* are intended to be used to locate words that need to be added to the provided dictionary and vocabular resources. If a word is not found, it should be added to these resource files (see Section 3.3).

First off, annotations in each interval of the transcribed file to be processed need to be exported to a separate file. With the appropriate audio and TextGrid files for talker 1 transcriptions (e.g. *s101alf-1.wav* and *s101alf-1.TextGrid*), run the Praat code file *1a extract tier annotations.praat*. Note that to run this code correctly, the current directory (*dir1*) and the SPPAS directory (*dir2*) need to be adjusted. This code will find all **-1.wav* and **-1.TextGrid* file pairs and export the annotations in each utterance into a newly created file labeled *annotations-...-1.txt*.

2.2 Identifying missing characters

In order to find Chinese characters or Latin-script words that cannot be found in the SPPAS vocabulary and dictionary files, you will need to run the code named *1b extract missing characters.r* using R software. This code was written in R because it makes use of more efficient programmatic functions that make the checking of annotations against a text dictionary much faster than coding with Praat. As with the first Praat code, the current directory (*main.dir*) and the SPPAS directory (*sppas.dir*) need to be adjusted according to your system.

Once you run the R code, a new file named *missing-chars-...-1.txt* will be created. This file contains all single Chinese characters or Latin-script letter sequences that cannot be found in the SPPAS vocabulary lists and dictionaries. Note that letter sequences such as *xxx* and *[=laughs]* are meaningless from the viewpoint of phonetic transcription, and so any intervals in the transcription files containing these letter sequences in their labels should be altered such that they are no longer present. One method for correcting this is going back into the TextGrid file and taking out the unintelligible or non-linguistic sequence of sound in the signal by making interval boundaries around cannot noise that cannot be processed.

2.3. Updating dictionaries

If a new word needs to be added to the SPPAS dictionary, the following files need to be updated:

- *yue.vocab* (*SPPAS > resources > vocab*)
- *yue_chars.vocab* (*SPPAS > resources > vocab*)
- *yue-monosyllabic.dict* (*SPPAS > resources > dict*)
- *yue.polysyllabic.dict* (*SPPAS > resources > dict*)
- *canto-pronunciation-dict.csv* (*SCOLAR-tools > ref-files*)

Shortcut files in the *1-check-dictionaries* (provided that they have been updated) can be helpful in quickly accessing these files when updating the dictionary.

Vocabulary files are the files that SPPAS uses to parse utterance transcriptions in each annotated interval into individual words or syllables. For **.vocab* files, only a single row for a new dictionary entry needs to be appended. For the *yue.vocab* (the vocabulary list for polysyllabic Cantonese words), new entries can be added to be end of the list. For the *yue_chars.vocab*, new monosyllabic Chinese character entries (or for

loanwords, entries for the entire word) can be appended in additional rows just above the fifth-to-last row (*!Enter*). Note that word and character entries cannot contain punctuation marks, apostrophes (’), or accented letters (e.g. *é*, *ü*, õ), and letter capitalization is essentially meaningless. Below are some examples for vocabulary list entries:

yue.vocab	*yue_chars.vocab*
o 唔 okay	o
saa1saa4	唔
c 出口	okay
snapchat	saa1
嘅 3	saa4
pet1pet1	c
welcome	出
pink	口
lettuce	snapchat
	嘅 3
	pet1

Note that in many cases, it will be unnecessary to add new characters or letter sequences to the *yue_chars.vocab* file, as this file already contains many of the elements that form dictionary entries in the *yue.vocab* word list. The *missing-chars…-1.txt* files are helpful in determining which words/characters can be found in each list. Words that involve repetitions of novel syllables require only one syllable entry in the *yue_chars.vocab* list (e.g. *pet1* for *pet1pet1*). Borrowings from non-Chinese languages can be broken up into individual words and syllables, but we have decided to leave them intact as the parsing of words and syllables along orthographic lines leads to unnecessary complications (e.g. orthographic vowels in English do not always denote vowels with actual phonetic value, e.g. *-e* in *s**e**ll* [s**ɛ**l], *s**ee**k* [s**i**k], *nic**e*** [naɪs], *fianc**ée*** [fi.an.s**eɪ**], etc.).

Dictionary (**.dict*) files can be updated by creating a new entries (one row each) for all possible phonetic variants of a given word or syllable. Contained within each entry row is the orthographic representation of the word/syllable, followed by a single space, then a pair of an opening and closing square brackets [], then another single space, and then finally a sequence of phonetic symbols in SPPAS’s own phonetic alphabet (**SAMPA** = **S**peech **A**ssessment **M**ethods of **P**honetic **A**lphabet), where each sound is separated by a single space. A more-or-less complete table of SAMPA to IPA correspondences are provided in Appendix 1. (Additional SAMPA symbols used can be found in the reference file *SPPAS-transcription-key.txt* in *SCOLAR-tools > ref-files*.) Note that in these transcriptions, linguistic tone is excluded entirely, as tone cannot be analyzed in terms of phonetic segments.

If you wish to add multiple phonetic variants for a given word/character, a number representing the order of the pronunciation variant (2 or above) set within brackets/parentheses (e.g. (2), (3), (4), etc.) should be included immediately after the orthographic item in the row. The presence of multiple phonetic variants in the dictionary is useful in cases where: (1) a given character has multiple canonical pronunciations, as in 樂, which can be pronounced as [lɔk^{6}] or [ŋɔk^{6}], or (2) a given word/syllable has multiple pronunciations in free variation, as in 你哋 [nei^{5}.tei^{6}]~[lei^{5}.tei^{6}], 啱啱 [aːm^{1}.aːm^{1}]~[ŋaːm^{1}.ŋaːm^{1}], etc. Note that, wherever possible, attempts were made to ensure that at least one Hong Kong English pronunciation variant was included among the list of possible pronunciations of lexical borrowings from English, and additional

variants were added if a particular pronunciation of such words in a recording deviated significantly from both standard as well as Hong Kong English forms. The following are some examples of novel dictionary entries in the corpus:

Entries in the file *yue-polysyllabic.dict*:	IPA equivalent
o 唔 okay [] o u: m o u: k_h e i:	[ou.m.ou.kʰei]
saa1saa4 [] s a: s a:	[saː.saː]
c 出口 [] s i: ts_h 8 t h 6 u:	[si.tsʰɐt.hɐu]
snapchat [] s n E: p tS_h E: t	[snɛp.tɕʰɛt]
嘅 3 [] k E:	[kɛ]
pet1pet1 [] p_h E: t p_h E: t	[pʰɛt.pʰɛt]
welcome [] w E: l k_h 6 m	[wɛl.kʰɐm]
welcome(2) [] w E: u: k_h 6 m	[wɛu.kʰɐm]
pink [] p_h I N	[pʰɪŋ]
pink(2) [] p_h E: N k_h	[pʰɛŋkʰ]
lettuce [] l E: t @ s	[lɛ.təs]
lettuce(2) [] l E: t_h 6 s	[lɛ.tʰɐs]

The last reference file that requires ongoing management is the file named *canto-pronunciation-dict.csv*, which is a comma-separated-values file that contains phonetic citation forms in both IPA and Jyutping romanization for each item in both the word-based and syllable-based dictionaries. This file will be an important component of the procedure in Section 4. In the *canto-pronunciation-dict.csv* file, you'll find entries for each syllable and word found in the vocabulary files, with each entry having a unique orthographic form and data fields being separated by commas. This file can be edited using a general text editor. Note that when editing this file, you must ensure that each row contains the correct number of data fields (6) and the correct number of separating commas (5). For all IPA and Jyutping transcriptions, syllables are separated by a full stop/period (.). The format of each entry is as follows:

> *Orthography*, *IPA citation form (with tones)*, *IPA citation form (tones omitted)*, *Jyutping citation form (with tones)*, *Jyutping citation form (tones omitted)*, *Tone sequence.*

For each of the transcription fields (fields 2-6), multiple citation-form variants are indicated with a set notation { *x* | *y* | *z* | … }, where *x*, *y*, *z*, … are possible citation forms. Note that only citation forms should be entered here, rather than possible non-standard variants. Thus, the entry for the word 我 only shows the single citation form *ŋɔ5* rather than two phonetic variants {ŋɔ5|ɔ5}, {ngo5|o5}, etc. Below are some examples of correct pronunciation dictionary entries (note that in IPA representation, it is preferred that syllabic nasals be represented using the syllabic diacritic below the nasal sound symbol, as in the first example):

> o 唔 okay,ou3.m̩4.ou3.kʰei1,ou.m̩.ou.kʰei,ou3.m4.ou3.kei1,ou.m.ou.kei,3.4.3.1
>
> saa1saa4,saː1.saː4,saː.saː,saa1.saa4,saa.saa,1.4
>
> c 出口, si1.tsɐt7.hɐu2 ,si.tsɐt.hɐu,si1.ceot7.hau,si.ceot.hau,1.7.2

snapchat,snɛp7.tɕʰɛt9,snɛp.tɕʰɛt,snep7.cet9,snep.cet,7.9

嘅 3,kɛ3,kɛ,ge3,ge,3

pet1pet1,pʰɛt7.pʰɛt7,pʰɛt.pʰɛt,pet7.pet7,pet.pet,7.7

welcome,wɛu1.kʰɐm6,wɛu.kʰɐm,weu1.kam6,weu.kam,1.6

pink,pʰɪŋ7,pʰɪŋ,ping7,ping,7

lettuce,lɛ1.tʰɐs9,lɛ.tʰɐs,le1.tas9,le.tas,1.9

In the case of non-Cantonese words, I have undergone the practice of transcribing citation forms as they would be pronounced by Hong Kong Cantonese speakers. This is broadly the case for English words and phrases. In such situations, the IPA citation forms more closely reflect the sound patterns of Hong Kong English speech, and the corresponding Jyutping romanizations are the nearest approximations of these sound representations using the Jyutping transcription system. A table of IPA to pseudo-Jyutping representations of English sounds not found in Cantonese are presented in Appendix 2.

3. Forced Alignment

3.1. Preparing word- and syllable-based analysis files

Before running the actual forced alignment procedure, the annotation files need to be separated into two separate analyses: (1) a word-based analysis that takes vocabulary items from *yue.vocab* and phonetic transcriptions from *yue-polysyllabic.dict*, and (2) a syllable-based analysis that takes vocabulary items from *yue_chars.vocab* and the corresponding phonetic dictionary *yue-monosyllabic.dict*. To do this, move the audio and TextGrid files from the *1-check-dictionaries* folder into the second folder *2-forced-alingment*, and then run the two Praat codes *1a add-IPA-tier-and-add-pause-marks.praat* and *1b split-into-syllable-and-word-file-sets.praat*. Note that you'll need to run each of these codes once for each recording you are processing, providing a new filename in the field *filename* at the top of the code.

The first Praat code will locate all utterance intervals in the input TextGrid file and make a new tier above the original tier in which all utterance intervals have been labeled *IPU_n* (where n is the order number of the interval), and all remaining empty intervals are labeled with the symbol #. This step is necessary for SPPAS to locate the intervals needing to be segmented. The annotation file is replaced with this new file, and a back-up version of the original TextGrid file is saved to the filename **-1-backup.TextGrid*.

The second code duplicates both the audio and the new TextGrid file into a pairs of word- and syllable-based analysis files (named **-1-words.wav*/*TextGrid* and **-1-syllables.wav*/*TextGrid*). Since we are making transcriptions for all words and for all syllables contained within those words, we need to run the analysis with both dictionaries to achieve the best lexical identifications. Ultimately, the phonetic transcription we will go with depends on what the output of the word-based analysis spits out, since this analysis provides the best upper-level (lexical) contextual information for the forced alignment analysis.

3.2. Performing forced alignment

To do the forced alignment, open SPPAS using the shortcut provided (with the adjusted shortcut path). Right when you open the program, a GUI (**G**raphical **U**ser **I**nterface) appears. At first, we need to perform the text normalization and phonetization procedures in SPPAS, whereby the orthographic transcriptions are parsed according to identifiable word or syllable items found in the vocabulary files (text normalization = tokenization) and the set of all possible phonetic variants are assigned to those items using the corresponding dictionary files (phonetization).

To perform the text normalization and the phonetization, we will need to do the word- and syllable-based analyses separately. For the word-based, add the word-analysis files by pressing the **Add files** button in SPPAS window, navigating to the *2-forced-alignment* folder, and selecting the **.wav* file that corresponds to the word-based analysis (e.g. *s101alf-1-words.wav*). SPPAS should load in both the audio file and its corresponding TextGrid file. On the right side of the menu, click on the **Annotate** button, and then scroll down to the **Text Normalization** and **Phonetization** options. Before proceeding, it's best to enlarge the width of the righthand pane by clicking on the vertical dividing line and dragging it towards the left (at least two-thirds of the way over). Doing so will enable to find the **Link/Unlink language selection** option on the right side (the small icon of chain links). Click it so that the language selections are unlinked (a yellow line appears in the middle of the chain icon). Then enable the Text Normalization and Phonetization options by clicking the red switch beside them so that they turn green. Next, in the pull-down menus on the right side, select the **yue** option for Text Normalization and select the **yue-polysyllabic** option for Phonetization. Then finally, scroll down to the bottom of the menu and click the **Perform annotations** button. This process should be rather brief, and the end of it, a procedure outcome report will appear. Close the report window, and check your file directory to see that three files have been created: **-1-words-token.TextGrid*, **-1-words-phon-TextGrid*, and **-1-words-merge.TextGrid*. The *token* file contains the segmentation of utterance annotations into individual words, and the *phon* file contains the same intervals but with all possible phonetic candidates assigned to each word token. The *merge* file is the merged file that contains information from both files as two parallel tiers.

After the word-based files have been processed, remove them from SPPAS by selecting on the entire folder (the loaded directory path on the left side of the menu) and then clicking on the **Remove** button. **Important:** Do not click the **Delete** button, as this will not just remove the files from SPPAS but it will also delete the actual files entirely from your computer! Now load the syllable-based files as before (click on **Add files**, navigate to **2-forced-alignment**, and select **-1-syllables.wav*. Once the **-1-syllables.wav*/*TextGrid* files have been loaded, readjust the Text Normalization setting so that **yue_chars** is selected and change the Phonetization setting so that **yue-monosyllabic** is selected. Then scroll down again and click **Perform annotations**. The output should be the files **-1-syllables-token.TextGrid*, **-1-syllables-phon.TextGrid*, and **-1-syllables-merge.TextGrid.*

The final (and most time-consuming) step in the forced-alignment procedure is running the actual Alignment operation on each of the files. These files can be aligned in a single batch, but SPPAS first needs to be closed and reopened (for reasons I'm not sure about myself). To immediately close SPPAS, find the power/shutdown button in the upper left corner of the SPPAS window and click it. Then reopen SPPAS using the handy SPPAS shortcut file.

Once you have reopened SPPAS, load both the **-1-words.wav* file and the **-1-syllables.wav* file using the **Add file** button. To select multiple files, simply click and drag (if files are consecutive), or click each file individually while holding down the Ctrl key (in Windows) or the Command ⌘ key (in MacOS), and then click **Open**. The set of *token*, *phon*, and *merge* annotation files for both the word- and syllable-based

files should appear on the left (10 files altogether). On the right side of the window, click on **Annotate** and once again expand the window by dragging the vertical pane divider towards the left. Then click again on the chain-link icon so that language selections are unlinked. Scroll down to the **Alignment** option and enable it by click on the red switch button. Then click on **Configure…**, and in the configuration window that appears, check the option **Perform basic alignment if the aligner fails**, and press OK (this step ensures that at least some forced alignment is given in the event that SPPAS runs into problems). Finally, set the pull-down menu for Alignment to **yue**, which uses the Cantonese phonetic-alignment model for our analysis, and the click **Perform annotations**. After this, you might be waiting for quite a while, depending on how much processing power your computer has and how long the audio files are. The procedure will be finished when a new **Procedure outcome report** window pops up. When this step has been completed, you can close SPPAS entirely. The files in the *2-forced-alignment* directory should appear almost as before, except that the two **-1-merge.TextGrid* files have been updated to contain new phonetic transcriptions and two new phonetic alignment files have been created (**-1-syllables-palign.TextGrid* and **-1-words-palign.TextGrid*). The important files here are actually the *merge.TextGrid* files, as they are files used as input in the next section.

3.3. Combining word- and syllable-based phonetic transcriptions

Once your files have been accurately aligned, you should have two merge files containing relatively accurate phonetic transcriptions in the word-based and syllable-based analysis. However, these transcriptions are still in the SAMPA format. Use the code *3 convert syllables or words - SAMPA to IPA.praat*, which takes the *words-merge* and *syllables-merge* files and converts them into IPA annotation files **-words-transcribed .TextGrid* and **-syllables-transcribed.TextGrid*, respectively. This code needs to be run two times, once with the **analysis.type** set to *words*, and another time with the **analysis.type** set to *syllables*. If the code crashes in the middle of running, it's likely that there was an issue with the forced alignment. In general, problems with forced alignment are due to one of the three following issues:

- The SPPAS vocabulary or dictionary files are missing the word or syllable item that appears in the orthographic transcription. This is often the case when uninterpretable sequences, such as *xxx* or *[=laughs]* has not been removed from the original transcription file.
- The entries in the pronunciation dictionaries contain errors, such as the wrong symbols or incorrect formatting.
- The phonetic symbol used in the phonetic dictionaries does not appear in the phonetic reference key files: *SPPAS-transcription-key.txt* and *diphthong-reference.txt* (both located in *SCOLAR-tools/ref-files*).

In general, the best way to fix such issues is to readjust the original transcription files and/or the corresponding dictionary files such that the error has been fixed. If a new phonetic symbol has been introduced (e.g. the SAMPA symbol *G* for the IPA sound [ɣ]), you will need to add this sound to the *SPPAS-transcription-key.txt* file located in the folder *SCOLAR-tools/ref-files*.

Once you have successfully generated the *transcribed* files (i.e. the phonetic alignment annotations converted into standard IPA), you will need to put the files from the two analyses together using the Praat code *4 combine transcription files.praat*. Run this code and you'll find a newly generated file named **-1-transcribed.TextGrid,* which combines tiers of the original utterance transcriptions with the orthographic and actual IPA representation tiers for both word and syllable levels, and additionally with the forced-alignment result from the word-based analysis in the tiers labeled *PHON-Phoneme* and *PHON-Phon*. The

PHON-Phoneme tier combines vowel segments within a syllable interval into a single diphthong, whereas the *PHON-Phon* tier separates such vowels into their own individual segmental intervals. In the ultimate corpus, we will draw from the phonemic (phonologically-driven) analysis rather than the strict phonetic analysis. When you have extracted the **-1-transcribed.TextGrid* file, you can move this file into the next subdirectory *3-add-canonical-forms*, delete any files that are no longer needed, and proceed to the next step. Note that the audio file (**.wav*) is no longer needed for future steps and can be deleted. However, I prefer to store the original transcription files (*-1.TextGrid and *-1-backup.TextGrid) with the primary audio files in case the forced-alignment procedure needs to be performed again.

4. Adding Citation Forms

4.1. Adding canonical phonetic forms

This section explains how to add phonetic citation forms in both Jyutping (粵拼) romanization and IPA transcription to the transcription files (*Actual IPA*). This step is simple as long as the reference file for citation forms *canto-pronunciation-dict.csv* is correctly formatted and contains entries for each word and syllable that occurs in the transcription file being processed. To apply the citation forms to your file, open and run the Praat code *1 add canonical forms.praat* (using the appropriate file directories). The output of this code is an annotation file named **-1-done.TextGrid*, which contains a large number of 16 interval annotation tiers. Move this file to the next directory *4-extract-data-files* for the next step. The **-1-transcribed.TextGrid* file can be deleted.

5. Extracting Data Files

5.1. Extracting transcript files

At this point, we are able to extract the orthographic transcription of our recording and store it with our corpus data. To do so, run the Praat code *1 extract-transcript-files.praat*. The code will all annotations in the Utterance (top) tier of the **-1done.TextGrid* file and print them into a single transcript file located in the raw data folder *SCOLAR-tools/corpus-data/raw*. In this file, each row states the offset time (in seconds) of each interval along with the annotation that occurred within that interval. For silent intervals, this annotation is marked as <*SIL*> (silence). Note that the onset of each row's interval is identical to the offset time of the interval immediately preceding it (except for the first interval, for which the onset time is 0.000 seconds). All transcript files have the file extension **.transcript*, and these files can be viewed using a simply-text editor.

5.2. Separating TextGrid files for data extraction

Before extracting other data files, I've included a step that again separates the annotations into word-based and syllable-based analyses. This makes the files easier to read and easier to work with. Separate the **-1-done.TextGrid* file by running the Praat code *2 split-into-two-TextGrids*. This code will create one annotation file for each of the two analyses (**-1-words.TextGrid* and **-1-sylls.TextGrid*). At this point, it is wise to save copies of these files along with the original files in the subdirectory under *audio-files*, as well as a set of copies uploaded to the corpus server (under *recordings/corpus*).

5.3. Extracting the rest of the data files

This step involves the extraction of files that lists the times of each word, syllable, phoneme, phone, and tone (if annotated) the recording at hand. It is advised that for Group 2 and 3 talkers, this step is put off until phonetic transcriptions from forced alignment have been checked for accuracy.

To extract the data files for words, syllables, phonemes, phones, and tones, make sure that the **-1-done.TextGrid*, **-1-words.TextGrid*, and **-1-sylls.TextGrid* files are located in the *4-extract-data-files* directory, and then run each of the following Praat codes:

- *3a extract-data-into-**words**-file.praat*
  Extracts the offset times and labels for each word interval, with the following information printed in the second column: *orthography*; *Jyutping*; *citation IPA*; *actual IPA*; *citation tone*; *actual tone* (In the IPA transcriptions, phonetic segments are separated by a single space, and syllable boundaries are separated by a full stop/period).
- *3b extract-data-into-**syllables**-file.praat*
  Extracts the offset times and labels for each syllable interval, with the following information printed in the second column: *orthography*; *Jyutping*; *citation IPA*; *actual IPA*; *citation tone*; *actual tone* (In the IPA transcriptions, phonetic segments are separated by a single space).
- *4a extract-data-into-**phonemes**-file.praat*
  Extracts the offset times of all phonemes, along with their phonetic identity in the second column. (Vowels within diphthongs are combined units.)
- *4b extract-data-into-**phones**-file.praat*
  Extracts the offset times of all phones, along with their phonetic identity in the second column. (Vowels within diphthongs are separate segments.)
- *4c extract-data-into-**tones**-file.praat*
  Extracts the offset times of all tones, along with their tone identity in the second column. (Note: This step is completely useless unless the actual tone data has been transcribed in the **-1-done.TextGrid* file.)

After the needed data files have been extracted, the annotation files (**.TextGrid*) can be removed and the raw data files (**.words*, **.syllables*, **.phonemes*, **.phones*, and **.tones*) can be moved to the raw data *folder: SCOLAR-tools/corpus-data/raw*.

6. Preparing Files for CLAN Tagging

6.1. Generating files for CLAN (Group 1)

To prepare annotation files that are ready for tagging in the CLAN system, the **-1-done.TextGrid*, **-1-words.TextGrid*, and **-1-sylls.TextGrid* files should be placed in the folder assigned to the subject within the *SCOLAR-tools/audio-files* directory. For recordings that contained two talkers (e.g. *cl* task recordings), ensure that the **-2.TextGrid* file, which contains only a single tier for annotations of utterances from the second talker in the recording, are found in the same subject-specific audio recording directory.

Next, in the *5-prepare-CLAN-files* folder, open either the code *1a generate-CLAN-TextGrid-Grp1-one-speaker.praat* (for one-talker recordings) or the code *1b generate-CLAN-TextGrid-Grp1-two-speakers.praat* (for two-talker/dyad recordings). The resulting file (**-CLAN.TextGrid*) should appear in the *5-prepare-CLAN-files* directory. These files are the format to be used to tag using the CLAN system for uploading to the CHILDES database. Note that the suffix of the resulting file is the same, no matter whether the recording had one or two talkers. However, for dyad recordings, the CLAN-ready file will contain nearly twice as many annotation tiers since there are two talkers. Once the file is ready, it can be uploaded to the shared project server for CLAN tagging.

6.2. Generating files for CLAN (Groups 2 and 3)

For the files for Groups 2 and 3, there is a necessary intervening step of checking the forced-alignment result manually, since the phonetic outputs of the talkers in these groups are not always accurately transcribed by the Cantonese phonetic model employed by SPPAS. Thus, additional code has been provided to generate files both for this checking stage as well as for reintegration into a format ready for CLAN tagging.

To progress a Group 2 or Group 3 file, select one of the two codes named *2a generate-CLAN-TextGrid-Grp2+3-one-speaker.praat* and *2b generate-CLAN-TextGrid-Grp2+3-two-speakers.praat*, according to the recording task (single or dual speaker recordings). After you run the code, a file named **-CLAN-unchecked.TextGrid* will be generated in the talker-specific audio file directory. This file will be used later down the line to reintegrate transcription corrects with the original pre-correction file.

After running either of these two codes, run the Praat code *3 prepare-checking-file-Grp2+3.praat* with the appropriate subject information. The code will produce a file labeled **-CLAN-**to-check**-TextGrid*, which should be given to assistants helping with transcription checking. When this file's transcriptions have been checked and corrected, it should be renamed to fit the naming format **-CLAN-**checked**.TextGrid*.

Lastly, the checked file needs to be reintegrated with the original forced-alignment result using one of the two Praat codes *4a reintegrate-checked-PHON-tier-Grp2+3-one-speaker.praat* and *4b reintegrate-checked-PHON-tier-Grp2+3-two-speakers.praat*. As before, use the code that is appropriate for the number of talkers transcribed in the recording. This code will build a new CLAN-ready file with word- and syllable-based analysis intervals that are based on the corrected transcription times in the corrected transcription file (**-CLAN-checked.TextGrid*). Additionally, the original phonetic transcription produced by forced alignment before correction will be found in the twelfth tier (labeled *PHON-Phoneme-PreCor-1*). The output file (**-CLAN.TextGrid*) should now be ready for CLAN tagging and then uploading onto the CHILDES database.

7. Appendices

7.1 Appendix 1: SAMPA to IPA correspondences

Tables below contain correspondences between HK Cantonese sounds in standard IPA and the SAMPA symbols used to represent those sounds in the SPPAS dictionary files. Jyutping representations are given in the third column for reference. Note that in the SAMPA, contrasts are made between lowercase and uppercase letter symbols, and each distinctive sound segment is separated by one single-space character. Rows highlighted in light grey describe sounds that occur in free variation with the sound in the row directly above them.

Consonants:

SAMPA	IPA	Jyutping
p	p, p̄	*b-*, *-p*
p_h	pʰ	*p-*
m	m, m̩	*m-*, *-m*
f	f	*f-*
t	t, t̄	*d-*, *-t*
t_h	tʰ	*t-*
ts	ts	*z-*
ts_h	tsʰ	*c-*
s	s	*s-*
n	n	*n-*, *-n*
l	l	*l-*
tS	tɕ	*z-*
tS_h	tɕʰ	*c-*
S	ɕ	*s-*
j	j	*j-*
k	k, k̄	*g-*, *-k*
k_h	kʰ	*k-*
N	ŋ, ŋ̄	*ng-*, *-ng*
w	w	*w-*
k_w	kʷ	*gw-*
k_h_w	kʷʰ	*kw-*
h	h	*h-*
?	ʔ	(n/a)

Vowels:

SAMPA	IPA	Jyutping
i:	i	i
i: u:	iu	iu
I	ɪ	i (+ng/k)
e	e	i (+ng/k)
e i:	ei	ei
I i:	ɪi	ei
E:	ɛ	e
E: u:	ɛu	eu
a:	aː	aa
a: i:	aːi	aai
a: u:	aːu	aau
6	ɐ	a
6 i:	ɐi	ai
6 u:	ɐu	au
u:	u	u
u: i:	ui	ui
U	ʊ	u (+ng/k)
o	o	u (+ng/k)
O: i:	ɔi	oi
o: u:	ou	ou
U u:	ʊu	ou
y:	y	yu
9:	œ	oe
8	ɵ	eo
8 y:	ɵy	eoi
@	ə	(n/a)

7.2 Appendix 2: English IPA to pseudo-Jyutping correspondences

Table below shows correspondences between the IPA representations of English sounds (or sound sequences) that do not occur in Cantonese and the pseudo-Jyutping system that I have devised in order to supply annotation files with canonical Jyutping sound representations.

IPA	**pseudo-Jyutping**	**Example words:**		
[b]	b- -p	*both* *ribeye*	[pouf^{7}] [ɹɪp^{7}.aːi^{6}]	*bouf7* *rip7.aai6*
[v]	v- -f	*eleven* *microwave*	[i^{3}.lɛ^{1}.vɐn^{6}] [maːi^{1}.kʰɹou^{3}.weifᵊ]	*i3.le1.van6* *maai1.krou3.weifᵊ*
[θ]	f	*three* *bath*	[fɹi^{1}] [pɛf^{7}]	*fri1* *bef7*
[ð]	d	*brother* *gathering*	[pɹaː^{1}.təɹ^{6}] [kæ^{1}.tɐ^{3}.ɹɪŋ^{6}]	*braa1.dar6* *ge1.da6.ring6*
[d]	d- -t	*dolphin* *record*	[tɔ^{1}.fɪn^{6}] [ɹɛk^{7}.kʰɔtᵊ]	*do1.fin6* *rek7.kotᵊ*
[sp]	sb-	*splash* *experiment*	[splɛɐ^{7}] [ɛk^{8}.spɛɹ^{1}.əɹ^{3}.mɛn^{6}]	*sbles7* *ek8.sber1.ar6.men6*
[st]	sd-	*story* *star*	[stɔ^{1}.ɹi^{6}] [staː^{1}]	*sdo1.ri6* *sdaa1*
[sk]	sg-	*skating* *business school*	[skei^{1}.tʰɪŋ^{6}] [pɪs^{7}.nɛs^{7}.sku^{1}]	*sgei1.ting6* *bis7.nes7.sgu1*
[z]	s	*fries* *organise/organize*	[fɹaːis^{7}] [ɔ^{1}.kɛ^{3}.naːisᵊ]	*fraais7* *o1.ge3.naaisᵊ*
[ɹ]	r	*pre-N* *forest*	[pʰɹi^{1}.ʔɛn^{1}] [fɔ^{1}.ɹɛsᵊ]	*pri1.en1* *fo1.resᵊ*
[tʃ]	c- -c	*check* *absolute pitch*	[tɐʰɛk^{7}] [ɛp^{7}.sou^{1}.lət^{7}.pʰitɐ^{7}]	*cek7* *ep7.sou1.leot7.pic7*
[dʒ]	z- -c	*giraffe* *orange*	[tɛɹ^{3}.ɹaːf^{7}] [ɔ^{1}.ɹɛntɐᵊ]	*zi3.raaf7* *o1.renc9*
[ʃ]	s	*shape* *sugar*	[ɕeip^{7}] [ɕʊk^{7}.kaː^{6}]	*seip7* *suk7.gaa6*
[g]	g- -k	*good job* *eggplant*	[kʊt^{7}.tɕɔpᵊ] [ɛk^{7}.pʰlaːntᵊ]	*gut7.zopᵊ* *ek7.plaantᵊ*
[æ]	e	*vaccine* *WeChat*	[vɛkᵊ.sin^{1}] [wi^{1}.tɕʰɛt^{7}]	*vekᵊ.sin1* *wi1.cet7*
[ɜ/ɚ]	oe	*earthquake* *turtle* *professor*	[ʔœf^{7}.kʷʰeikᵊ] [tʰœ^{1}.tʰou^{6}] [pʰɹou^{3}.fɛ^{1}.sœ^{6}]	*oef7.kweikᵊ* *toe1.tou6* *prou3.fe1.soe6*
[ʌ,ə]	a	*mother* *lettuce*	[maː^{1}.tɐ^{6}] [lɛ^{1}.tʰɐsᵊ]	*maa1.da6* *le1.tasᵊ*