NSC Transcription Notation – Summary

	TYPE	CONVENTION	EXAMPLE
1.	End of sentence**	<s/>	how's your mum **<c/>** dad and sis **<s/>** last june **<c/>** they went to korea **<c/>** right **<s/>** ****Please ignore punctuation for now**
	Comma**	<c/>	
2.	Numbers	full form	today is **first** jan **twenty eighteen** <s/> total is **eleven dollars eighteen cents** <s/> cool <s/>
	Symbols	full form	
3.	Titles	full form (unless abbreviated)	**doctor** li-ann met **mister** muhamad-nor-hisham <c/> **miss** (um) **madam** aarti and **professor** ang at this year's meet up <s/>
4.	Acronyms	underscore '_'	he **wanna** study in **NAFA** <c/> but **couldn't** get in <s/> so **he's** now **gonna** join **N_T_U** to study **L_M_S** <c/> linguistics i mean <s/>
	Word-like acronyms	full form	
	Contractions	as is (unless unabbreviated)	
5.	Discourse particles	square bracket '[...]'	**[oh]** he still wanna eat **[ah]** <s/> **!wah!** why he like that **[one]** <s/> damn greedy **[lah]** he <s/>
6.	Fillers	round bracket '(...)'	**(uh)** ya <c/> it's near H_S_S <s/> to get there <c/> **(um)** i think you can turn left and **(err)** actually turn right and walk straight <s/> you should reach in **(um)** five minutes <s/>
7.	Interjections	Exclamation mark '!...!'	**!walao!** why the hell do we need to study English [right]
8.	Paralinguistic Phenomena	(ppb) breath (ppc) cough (ppl) laugh (ppo) others	**(ppc)** (um) there was once i confidently strut down the catwalk <c/> as though i'm **(ppo)** naomi-campbell <s/> then i **(ppb)** trip over a stone and **(ppl)** fell flat on my face <s/> **(ppl)**
9.	Other languages	hashtag '#...#'	she went **#pasar malam#** to buy **#roti-john#** lucky **#muah-chee#** **#shiok#** else she'll be **#pek chek#** **See below for explanation.**

10.	Unclear words	<UNK>	i went to **<UNK>** <c/> (err) or was it (um) **<UNK>** <s/> i can't remember the name <s/>
11.	Incompletely uttered words	tilde '~'	[oh] i don't understand the **abbre~ abbrev~** (uh) abbreviation that they use nowadays <s/> **not to be used for truncated words eg. prof (as in professor), lit (as in literature)
12.	Personally identifying info	<P1>…</P1>	<P1>**Tom Tan Ah Kao**</P1>
		<P2>…</P2>	<P2>**Emily Tan**</P2>
13.	Hate speech/ Critique/Crude Language	<H1>…</H1>	<H1>**gays should be sent to concentration camps**</H1>
		<H2>…</H2>	<H2>!walao! he damn **gayshit** [sia]</H2>
14.	Expletives	<EX1>…</EX1>	this stupid <EX1>**apuneneh**</EX1> was like
		<EX2>…</EX2>	!walau! you <EX2>**cheebye** kia</EX2>
15.	Short pause (longer that 1000 ms)	<S>	
16.	Invalid	<Z>	
17.	Long-running Non-English utterances	<NEN>	[eh] speaking of which right okay **<NEN>** we'll just talk okay when I watch the Korean show [ah] <c/> the actor say **<NEN>** and I'm like !wah! damn starstruck [sia]. This is reserved only for languages besides Malay, Mandarin and Tamil eg. Hindi, Korean, Japanese etc. An example of foreign language codeswitching would "naneun maeil haggyoe ganda" in Korean. Instead of

			transcribing the full utterance, just use the tag <NEN>.
18.	Long-running Malay utterances	<malay></malay>	yesterday I wanted to go to the movies <malay>**tapi aku sakit**</malay>
19.	Long-running Mandarin/Tamil utterances	<mandarin>xx: yy</mandarin> Where xx is Chinese/Tamil characters, and yy is Hanyu Pinyin/Romanised Tamil	I'm like thinking why she ask me <mandarin>你吃饱了吗:**ni chi bao le ma**</mandarin> [dey] <tamil>**நல்ல:nalla**</tamil> movie [lah]
20.	Redacted	**	this stupid ** was like

**Ignore punctuation markers <s/> and <c/> for now.

Overview

Transcription is the transformation of verbal and audio materials into text. In this project, voice recordings of Singaporeans from all ethnicities and backgrounds are collected to enrich the corpus. The aim of the National Speech Corpus (NSC) is to increase the accuracy of speech recognition technology in recognizing and transcribing Singapore-accented English, as well as a resource for researchers doing language-related research.

Automatic speech recognition (ASR) technology may be used to increase the efficiency in transcribing the voice recordings. Due to the ASR's imperfect knowledge of human language, there are many features in human speech that machines are not able to handle well, such as rapid or unclear speech, variability (gender, accent, slangs, etc), specialized or uncommon words and so on.

The task of human transcribers is to manually verify the machine transcript and amend the errors according.

See below for NSC's specific transcription requirements:

1. Your main task as a transcriber is to transcribe **VERBATIM** (type what the person was saying exactly in the recording). The transcript should be an **ACCURATE** representation of the original conversation.
2. Speech recognition tool might be used to generate the transcripts of the recordings. The machine transcripts are full of inaccuracies (missing or out-of-context words) and also lack other information like notation symbols.
3. For the NSC transcription task, you need to rectify the incorrect words in the transcripts and add notational information to the transcript.
4. The end product of the transcript should i) be an accurate representation of the conversation in the recording, ii) meet the notation requirement of the NSC transcription guidelines

Transcription Conventions

Punctuations

- Boundaries must be marked at sentence level, down to the last phoneme.

- Conventions:

1.	End of sentence	**<s/>**	e.g. '.', '!', '?'
2.	Comma	**<c/>**	e.g. ','

- Example:
  i'm fine **<c/>** thanks **<s/>** how are you **<s/>** i haven't seen you in ages **<s/>** if you drop by jurong again **<c/>** don't forget to ring me up <s/> we can chill **<c/>** go shopping **<c/>** and catch up over a cup of coffee **<s/>**

Numbers and Symbols

- Numbers and symbols must be represented in words.

- Conventions:

1.	Numbers	**full form**	e.g. thirtieth, twenty eighteen
2.	Symbols	**full form**	e.g. dollars, percent, plus

- Example:
  their **tenth** anniversary concert will be held on **fourth** july **two zero one eight** <s/> around **three hundred fifty** students bought the tickets with a **twenty percent** discount <s/> [woah] that's like **minus ninety dollars** from the original price <s/>

Titles

- Titles must be represented in their full form (unless abbreviated in the audio).

- Conventions:

1.	Titles	**full form**	e.g. mister, missus, professor

- Example:
  yesterday <c/> i watched **doctor**-doolittle again on netflix <c/> and then met up **professor** james-tan and **missus** siti-maisarah at coffee-bean around noon <s/> what a coincidence <c/> **mister** low and **miss** chan happened to be there as well <s/> they were celebrating **mister** low's promotion to an **associate prof** <s/>

  (*note: in this case, *associate prof* is transcribed as such, since *professor* is being abbreviated to *prof* in the audio)

Abbreviations

- Abbreviations must be represented in their abbreviated form (unless unabbreviated in the audio).

- Conventions:

1.	Acronyms (individual letters)	**underscore '_'**	e.g. M_O_E, T_V, U_K, H_and_M
2.	Acronyms (word like)	**Full form**	Eg. NATO, MINDEF, UNESCO
3.	Contractions	**abbreviated form**	e.g. you're, wanna, cause

- Example:
  besides **NAPFA** <c/> i'm okay with national-service [lah] <s/> after **N_S** <c/> i literally just watch **T_V** all day <c/> every day <c/> while waiting to enter **N_T_U** <s/> [oh] i did travel to **U_S** before that <s/> i went new-york <c/> **N_Y_C** <s/> i visited times-square <s/> but **didn't** manage to go central-park **cause** too rushing <s/> i also went to los-angeles <c/> my cousins are staying in **L_A** <c/> so they showed me around their **U_C_L_A** campus <s/> **they're gonna** come to singapore next week <s/>

  (*note: in this case, *national-service*, *new-york* and *los-angeles* are transcribed as such, since they are unabbreviated in the audio)

Discourse Particles

- Discourse particles, or pragmatic markers which directs the flow of conversation without adding significant paraphrasable meaning to the discourse, must be written between <u>square brackets</u>.

- Conventions:

1.	Discourse particles	**square bracket '[...]'**	e.g. [oh], [oo], [lah], [eh], [liao], [lor], [sia], [one]

- Example:
  !woah**!** the place was freaking huge **[lah]** <s/> #siao# **[liao] [lor]** <c/> can #mati# **[sia]** <s/> w_t_h <c/> there's no way anyone can finish in one day **[man]** <c/> unless he some kind of superman **[ah]** <s/> **[eh]** wait <c/> i think my brother confirm can **[one]** <c/> cause he likes this kinda arty farty things <s/>

Fillers

- Fillers, or meaningless sounds that marks a pause or hesitation in speech, must be written between <u>round brackets</u>.

- Conventions:

1.	Fillers	**round bracket '(...)'**	e.g. (um), (uh), (err), (hm)

- Example:
  (um) last year [ah] <s/> where did i go last year <s/> **(err) (hm)** so hard to remember <c/> so old already [lah] me <s/> (ppl) **(um)** if i'm not wrong <c/> i went to **(err)** south-korea with my mum to celebrate my (ppb) twenty sixth birthday <s/> [eh] wait <c/> or was it twenty fifth <s/> **(err)** nineteen eighty nine <c/> now is twenty eighteen <s/> **(uh)** sorry [man] <c/> my maths damn #rabak# <s/> (ppl)

Interjections

- Interjections, or words added to a sentence to convey an emotion or a sentiment such as surprise, disgust, joy, excitement, or enthusiasm, must be written between 2 exclamation marks.

- Conventions:

1.	Interventions	**Exclamation marks '!...!'**	e.g !wow!, !aiyo!, !wah!

- Example:

!aiyo! <s/> why he like that [one] <s/> **!eeyer!** <s/> can he don't be so bad or not <s/> **!hais!** I feel so bad for him [sia] <s/>

Paralinguistic Phenomena

- Paralinguistic phenomena, or non-speech sounds like breathing, coughing and laughing, must be written between <u>round brackets</u>.

- Conventions:

1.	Paralinguistic phenomena	**(ppb)**	breathing
		(ppc)	coughing
		(ppl)	laughing
		(ppo)	Others (other non-linguistic noise from the main speaker eg. tsk)

- Example:
  (ppo) so he went home <c/> thinking that <c/> **(ppc)** sorry <c/> thinking that the long day was over <s/> then he was just gonna step into the house <c/> **(ppb)** his phone rang <s/> the hospital asked him to come back **(ppc)** cause of an emergency <s/> so poor thing he <s/> **(ppl)** (err) no choice [lor] <s/> life of a doctor <s/> **(ppb)**

Other languages

- Words from other languages must be written between <u>hashtags</u>.
- **DO NOT** confuse this with language tags below – the hashtags are used when you encounter words from

a) **Chinese dialects** (e.g. Hokkien, Cantonese, Teochew),

b) from **languages other than Malay, Mandarin and Tamil**, or
c) for words that can be considered **Singlish**: An expression is considered Singlish when the non-English word (including words originating from Malay, Mandarin and Tamil) is 1) adopted by other ethnicities, 2) and has widespread use in Singapore English.

- Conventions:

  1. Other languages **hashtag '#...#'** e.g. #bo pakai#, #kiasu#, #kueh-lapis#, #roti-prata#

- Example:
  singapore slang <s/> (hm) let me think <s/> first <c/> you confirm hear this before <c/> singaporeans so **#kiasu#** <s/> ya <c/> sometimes join queue also don't know for what <s/> (ppl) then there's **#shiok#** <s/> you can use it when you **#makan#** something delicious <s/> now even got **#shioklicious#** [lah] <s/> (ppl) then if you cannot **#tahan#** <c/> i mean cannot tolerate someone <c/> cause he so **#bodoh#** or something <c/> then you say (uh) **#buay tahan#** <s/> (um) what else <s/> (ppb) [oh] for food and drink <c> there's chicken-rice <c/> **#wanton-noodle#** <c/> **#nasi-lemak#** <c/> **#roti-prata#** <c/> **#ice-kacang#** and (um) **#teh-tarik#** <s/>

Unclear Words

- Unclear words, or words that transcribers are unsure of, must be <u>included</u> in the transcription.

- Conventions:

  1. Unclear words **<UNK>**

- Example:
  she said when she was there <c/> she visited the **<UNK>** <c/> or was it **<UNK>** <s/> (hm) i'm not sure what the name is <c/> but anyway <c/> i heard it's (err) a really pretty place <c/> it's like (uh) a go to place when you visit that city <c/> ya <s/>

To take note:

Words that are obviously misprounounced are to be treated as unclear words, and to be annotated with <UNK> as well – e.g. maternity pronounced as "merterterly" should be written as <UNK>.

Slight mispronounciation due to accent or slight verbal error are to be written as the proper, intended word – e.g. number pronounced as "nummer" should be written as "number" in the transcrit).

Incompletely Uttered Words

- Incomplete words must be written as verbatim.
- Don't confuse incompletely uttered words with abbreviations – words like prof (as in professor) or lit (as in literature) are considered full words and should be spelt as how they are being pronounced.

- Conventions:

  1. Incomplete words **tilde '~'** e.g. idiosync~, disambig~

- Example:
  mary is **ambix~ ambidextr~** (um) ambidextrous is it <s/> ya <c/> ambidextrous <c/> can use both right and left hands to write <s/> [oh] [oh] but right <c/> i heard that this type of people more likely to be **schiz~** (err) **~phrenic** <c/> what's that word <c/> schizophrenic [ah] <s/> ya <s/> o_m_g <c/> that's scary [siot] <s/>

Personally Identifying Information

- Personally identifying information must be tagged. Tier 1 is for information that directly identifies an individual and must be silenced/censored. Tier 2 is for information that does not, but can be combined to identify an individual. Context dependent.

- Conventions:
  1. Tier 1 Personal Information **<P1> </P1>**
     - eg. <P1>Ben Tan-Wei-Ming</P1>
  2. Tier 2 Personal Information **<P2> </P2>**
     - eg. <P2>Ben Tan</P2>

Hate Speech

- Hate speech must be tagged. Tier 1 is for hate speech that is either a) racially or religiously charged and/or b) abuses or threatens a marginalised group. Tier 2 is for hate speech that certain groups may find offensive, but is not harmful.

- Conventions:
  1. Tier 1 Hate Speech **<H1> </H1>**
     - eg. <H1>gays should be sent to concentration camps</H1>
  2. Tier 2 Hate Speech **<H2> </H2>**
     - eg. <H2>!walau! this is so gayshit [sia]</H2>

Expletives

- Expletives must be tagged. Tier 1 is for expletives that are racially or religiously charged, and must be silenced/censored. Tier 2 is for common swear words that some may find offensive, but are not harmful.
  1. Tier 1 Expletives **<EX1> </EX1>**
     - Eg. <EX1>fucking chink</EX1>
  2. Tier 2 Expletives **<EX2> </EX2>**
     - Eg. <EX2>ni nabeh chao chee bye</EX2>

Note: Expletive with widespread use and is not considered harmful and offensive by most people do not need to be tagged. This includes the most commonly used vulgar terms in Singapore English: "fuck" (from English), "他妈的" (from Mandarin), "cheebye" (from Hokkien), "kimak" (from Malay) and "pundeh" (form Tamil).

Short Pause

- Short pauses, or periods of silence that are longer that 1000 ms must be marked. Always an entire interval.
  1. Short pause **<S>**

Invalid part

- <Z> refers to discernable noise not from the primary speaker – i.e voice of the experimenter or crosstalk from the other speaker in the recording.
- Invalid parts of the speech must be marked. Always an entire interval. Invalid parts include:
  - Invalid speech
  - Noise
  - Non-primary speaker's voice
  - Sounds from the monitor/speaker
  - Continuous noise eg. long laughs, music, singing etc.

1. Invalid part **<Z>**

Run-on Non-English Utterances

- Long utterances that are not in English can be tagged as <NEN>.

- **Do not mistake for #...#**, this notation is purely for singular words or short phrases of other languages that can be transcribed immediately.
- Will be transcribed at a later phase.
- Example: ya **<NEN>** but basic like **<NEN>** basic **<NEN>**

Run-on Malay Utterances

- Long utterances in Malay but now be fully transcribed and tagged using <malay></malay>.
- **Do not mistake for #...#**, this notation is purely for singular words or short phrases of Malay that a) are adapted into Singapore English and b) can be transcribed immediately.
- Example: yesterday I wanted to go to the movies **<malay>**tapi aku sakit**</malay>**

Run-on Mandarin/Tamil Utterances

- Long utterances in Mandarin/Tamil will be fully transcribed and tagged using <mother tongue>xx:yyy</mother tongue>, where x is in Chinese/Tamil characters and Y is in Hanyu Pinyin/Romanised Tamil.
- Hanyu Pinyin is used to romanise Mandarin utterances but with slight modifications
    - No diacritics/tone marks (i.e. “tai yang” and not “tàiyáng” or “tai4 yang2”)
    - Each syllable should be spelt apart (i.e “xin jia po ren” and not “Xinjiapo” ren)
    - Words spelled with umlauts in hanyu pinyin (such as 女、旅) should be spelled as “v” and not “ü” (i.e. “nv ren” and not “nü ren”).

- **Do not mistake for #...#**, this notation is purely for singular words or short phrases of Mandarin or dialects that a) are adapted into Singapore English and b) can be transcribed immediately.
- Mandarin example: I’m like thinking why she ask me **<mandarin>**你吃饱了吗:ni chi bao le ma**</mandarin>**

- Tamil example: [dey] **<tamil>**நல்ல:nalla**</tamil>** movie [lah]

Redacted

- After the silencing of sensitive material earlier tagged as either P, H or EX, the text in which were marked will now be replaced with ** to indicate the presence of earlier content that has since been removed.

Additional Principles:

- Tagging Non-English toponyms (place names):
  - Leave country names alone. E.g. Uzbekistan, Mozambique
  - Tag everything else e.g. #Taipei#, #Monte Carlo#, #Beijing#, #Toa-Payoh#

- Hyphen in enclosing tags are optional
  - #bak-chor-mee# is same as #bak chor mee#

- If there are hyphens in the orthography, use the hyphen
  - E.g. anti-government

- Tag what you hear not what they say
  - E.g. Breadtalk vs Breaktalk

- Bound morphemes on tagged terms
  - E.g. C_C_As, O_R_Ded

Sensitive Data Overview

Type	Tier	Details	Examples
Personally Identifying Information	P1	Information that directly identifies an individual: a. Name b. Place of work/school c. Salary d. Place of living e. Biometric data e.g. NRIC, passport number etc. f. Health data e.g. STDs, mental health g. Contact details e.g. HP number, email, usernames	• <P1>Ben Tan-Wei-Ming</P1> <P1>Kevin Dingwei Khoo</P1> • <P1>Nur Atiqah Binte Mohd Hamid </P1> <P1>Atiqah Hamid</P1> • <P1>Govindran Veerasamy</P1> • I go to <P1>Mass Comm at Ngee Ann Poly, year 3<P1> • My insta username is <P1>rocketman23</P1> • Oh ya my IC number is very weird it's like <P1>G nine four eight three and something something I don't remember and the last letter is F</P1> • He disappeared for a while [lah] went to get treatment for <P1>H_I_V</P1>
	P2	Information that has some semblance of identification, but not completely (becomes identifiable with other data). Context Dependent.	• <P2>Ben Tan</P2> • <P2>Kevin Khoo</P2> • <P2>Nur Atiqah</P2> • <P2>ya I did early childhood in poly<P2/> • <P2>ya I live at Boon Lay near extension there</P2> • I mean I have <P2>depression</P2> and I have to go therapy at <P2>N_U_H</P2> every month [lah] so

Hate Speech/Critique	H1	Speech that is abusive and threatening of following topics: a. Racially charged b. Religiously charge c. Personal attacks d. Attacks against marginalised groups e. Extremely crude language	• <H1>that make me think a then I would think !wah! if the negro also can pick up</H1> • <H1>Someone should bring a sniper and shoot her</H1> • <H1>Malay people should stop breeding<H1/>
	H2	Speech that is offensive to some, including crude language. Context Dependent.	• <H2>Who masturbated to a peach</H2> • <H2>!walao! he damn gayshit [sia]</H2>
Expletives	EX1	Expletives that must be censored/silenced because they are: 1. Racially charged 2. Religiously charged	Nigger, Nigga, Apuneneh, Paki, Chink, Fag, Faggot, Negro, Bangla
	EX2	Need not be censored/ silenced: offensive but not in relation to a person or a marginalised group. Context Dependent.	Nabei, Cheebye kia, Cunt, Pok Kai Ham Ka Chan

Annex A: Doing good transcription

Overview

The speakers are recorded separately (each speaker with their individual tracks) for the NSC project.

- Issue of human transcribers:
    - Difficulty comprehending certain segments (especially unclear or short utterances) when editing one speaker at a time due to lack of full picture of the conversation.
    - Accept technology unquestionably without questioning the accuracy of machine transcription.
    - May, at times, transcribe phonetically or type what they think they hear
- Solution:
    - **Load the two tracks (Speaker A and Speaker B) in Audacity** (Go to the link to download the software - https://www.audacityteam.org/download/) **and listen to the complete recording** in order to understand the full conversation.

Observations and Errors

Automatic speech recognition technology may be used to reduce the efforts needed to transcribe the recordings. Here are the general mistakes in machine transcription:

- Insertion (Additional words in transcript not present in the recording)
- Subtraction (Missing words in the transcript present in the recording) and Substitution ("Mishearing" the conversation and producing wrong words that are similar sounding).
    - E.g of substitution: misrecognising "session" as "fashion" in the utterance "Session 5 Language Policy Interview"
    - Main errors to be manually corrected by humans.

Machine can generate sentences that are **gibberish** to human readers as it does not understand language:

Error: Go and ride and explain the Eiffel have deducted

Correct version: Go and explain that I have deducted

Context: Financial transaction enquiry

Error: All the more assault teck leng creatures

Correct version: Or the more exotic languages

Context: Language learning

Users who transcribe without considering the context of the conversation may **erroneously** type homophones or similar sounding words:

Error: Then what about **languish** teaching
Correct version: Then what about language teaching
Context: Interview on language usage

Error: Do you like came into this **cause** of yours
Context: Do you like came into this course of yours
Context: University-related conversation

Rule of thumb for doing good transcription

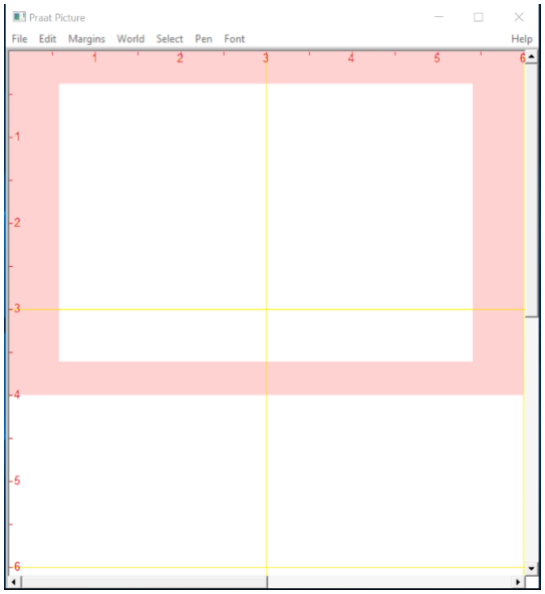
- Imperative for human transcribers to **identify incoherent sentences** when editing the transcripts
- **Transcribe or correct transcription errors with words that are appropriate to the context of the recording**s.
- The words in the transcript should not only align to the context, but the words in the two TextGrid files (representing Speaker A and B) should agree with each other
- A good transcriber should never
    - assume the machine transcription is fully accurate
    - transcribe according to their subjective interpretation of the utterance.

*IMPORTANT NOTICE: At all times, the human transcriber should consider **the context, topic and the interlocuters' perspectives from both sides of the recordings** when editing the transcripts

Annex B: Using Praat

Praat is a free software program that can be used for the analysis and reconstruction of acoustic speech signals. For this transcription project, it will be used to edit/ verify the processed transcripts. The *Praat* software can be downloaded here[1].

1. When *Praat* is launched, the two following windows will appear:



2. For this project, only the window '*Praat Objects*' (the one on the left) will be used. You may close the window *'Praat Picture'* since it is not required for the transcription process.

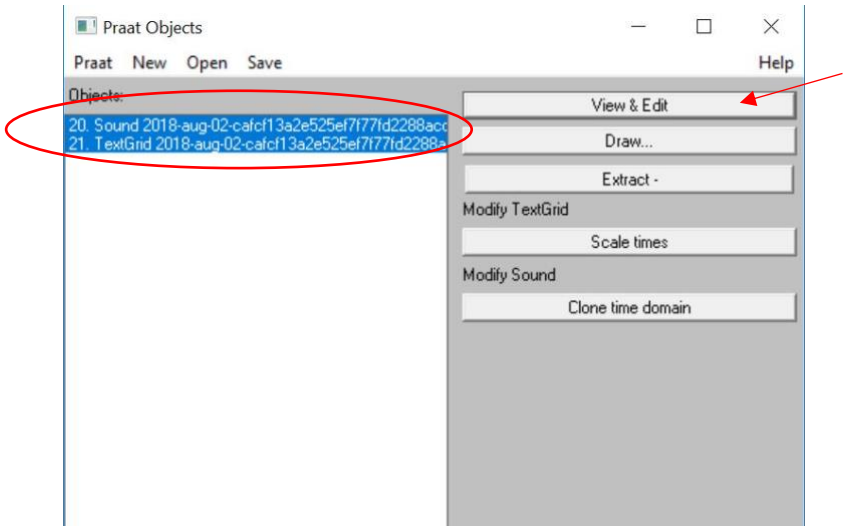
3. To open the *Soundfile* and *TextGrid* files, click <Open> then <Read from file…> in the *Praat Objects* window.

Soundfile: "(File ID).wav file"
TextGrid: "(File ID).TextGrid file"

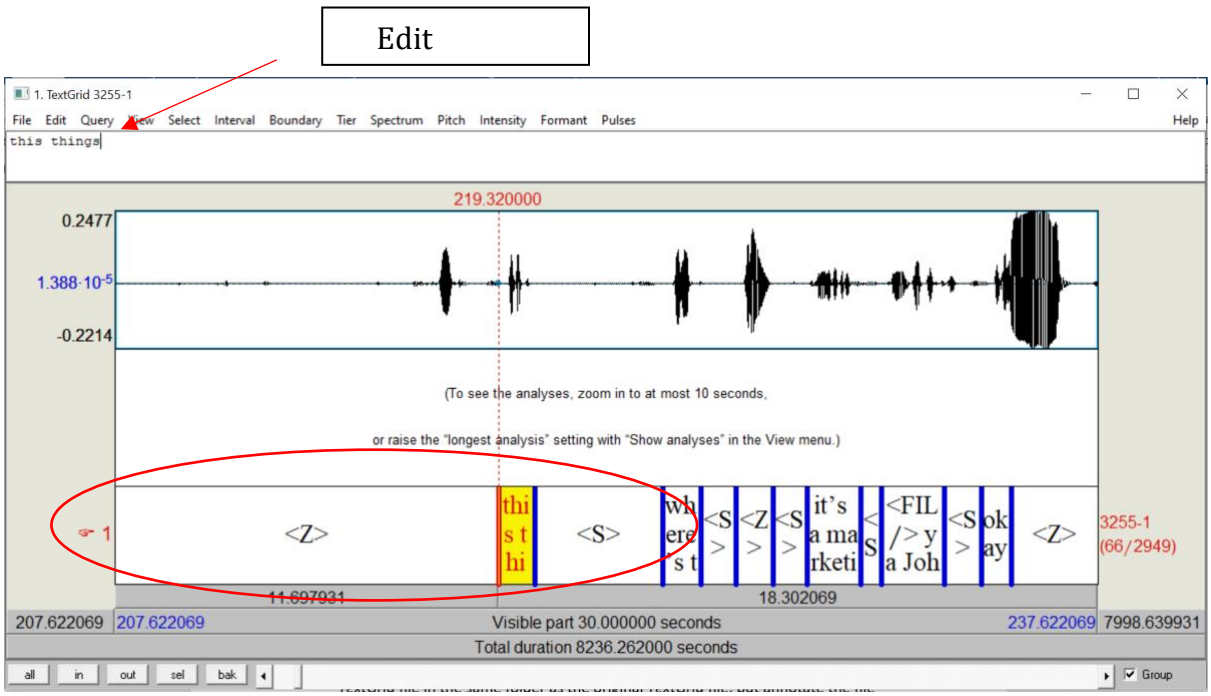
[1] http://www.fon.hum.uva.nl/praat/

1. Editing *TextGrid* files

1. To view and edit the transcripts, select both *Soundfile* and *TextGrid* files (*Ctrl+ left click* to select both items), then click <View & Edit> on the right-hand side of the window.



2. To edit the transcript, click on any of the segmented areas at the bottom. Press *Tab* to play the audio and then edit the transcript accordingly.



Note: **Do not** edit segments that are marked <S> or <Z> unless they contain speech data, and **do not** adjust the pre-defined boundaries.

After you have finished transcribing or correcting the scripts, you may save the TextGrid as text file, and remember to ensure that the file format is reflected in the name of the document – i.e it should have .TextGrid at the back of the name.

Annex C: Style Guide

1. **Spelling**
   - Use the spelling for UK English, not US English.
   - Eg. Colour instead of color, Harbour instead of Harbor

2. **Lowercase letters**
   - When typing out the transcription, please only use **lowercase** letters. Uppercase letters (capital letters) are only reserved for the following cases:

   a) Acronyms eg. I_M_D_A, M_O_E
   b) First person pronouns (I went to school, I just came back)
   c) Individual letters eg. when someone is spelling out something (I spelt it as O K R A) Note: for cases like these, you do not use underscore.

3. **Numbers**
   - Numbers should be SPELT out, not typed as digits
   - Eg. COVID nineteen

4. **Corporate Brand Names**
   - There are plenty of brand names that make use of a mix of words from English, foreign languages, individual letters, symbols even.
   - When you encounter these, please type the brand name as it is to be represented and **capitalize the first letter**.
   - DO NOT USE the annotations/tagging system as stated in the guidelines, it makes them look overly complicated.
   - This applies to both FOREIGN and LOCAL brands.
   - DO NOT HASHTAG even if they are foreign sounding.
   - Eg. Airbnb, Koolaid, My First Skool, UTorrent

5. **Noun Phrases with the word "and"**
   - Words that include the word "and" are typically shortened to use the ampersand "&" symbol.
   - Do not space the letters out, keep them compact to a single word.
   - Eg. F&B, R&D, A&E.

6. **Words with punctuation**
   - Words that are normally spelt with punctuation can be represented with the punctuation eg. O'clock, Ma'am.
   - Also applies for genitive case [possession aprostaphe] eg. Sara's dog, and hyphenated words eg, pre-requisite, ex-husband, mother-in-law

7. **Hybrid words**: When English rules are attached to non-English words
   - In Singapore English there is a tendency for speakers to **attach English grammatical rules** (Eg. Plural –s or -es, present participle –ing) to **non-English words**. These become eg. Makan-ing, Chee Kueh-s, tempeh –s, etc.
   - When you encounter cases like this, please **do not use hyphen**, but instead, **spell the word as if it were whole**, and **hashtag the entire term**.
   - It will therefore look like this: #makaning# #chee kuehs# #tempehs#
   - Note: For **CHINESE WORDS** that are hybrid, use Hashtag and leave the base Chinese word in Chinese character form eg. #课文 s#

8. **Language tags**
   - We have 3 types of language tags:
     - Hashtags (##),
     - Long Running Non-English tag (<NEN>) and
     - Language Tags for Malay, Mandarin and Tamil (<language></language>)
   - These 3 tags have very different uses. Please do not use them wrongly.

Hashtags:
Meant for singular words/short phrases from the following cases
a) **Chinese dialects** (e.g. Hokkien, Cantonese, Teochew),
b) from languages **other than** Malay, Mandarin and Tamil, or
c) for words that can be considered **Singlish**: An expression is considered Singlish when the non-English word (including words originating from Malay, Mandarin and Tamil) is 1) adopted by other ethnicities, 2) and has widespread use in Singapore English.

Long Running Non-English tag:
Meant for **long utterances** that are **not** in English, Malay, Mandarin or Tamil.

Malay, Mandarin or Tamil tags:
Meant for long utterances in Malay, Mandarin and Tamil that are to be fully transcribed according to the formats mentioned in the NSC Transcription Guidelines.

Annex D: list of approved particles, fillers and interjections

[discourse particles]	(fillers)	!interjections!
a'ah	err	aduh
ah	hmm	aiya
bah	mm	aiyo
dah	mmhmm	aiyoyo
dey	oo	alah
eh	ugh	alamak
ha	uh	aww
ho	uh huh	chey
hor	um	choy
jer		duh
kan		ee
lah		eeyer
leh		eww
loh		fuh
lor		hais
mah		haiya
meh		hanna
oh		hannor
one		hey
orh		huh
pe		oho
seh		oi
sia		oof
sial		oops

siol		ouch
tau		ow
what		siala
		wah
		wahseh
		wahpiang
		walao
		waliao
		waliew
		whew
		whoa
		woo
		woohoo
		wow
		yay

Annex D: Transcribing Chinese codeswitching utterances

1. Romanising Mandarin utterances

- Hanyu Pinyin is used to romanise Mandarin utterances but with some modifications:
    - No diacritics/tone marks (i.e. “tai yang” and not “tàiyáng” or “tai4 yang2”)
    - Each syllable should be spelt apart (i.e “xin jia po ren” and not “Xinjiapo” ren)
    - Words spelled with umlauts in hanyu pinyin (such as 女、旅) should be spelled as “v” and not “ü” (i.e. 女人 is “nv ren” and not “nü ren”).
- Format of transcribing Mandarin utterances will be written as <mandarin>xx:yyy</mandarin>, where xx is in Chinese characters and yyy is in modified hanyu pinyin.
    - <mandarin>我看完这部戏了了我非常喜欢他扮演的角色:wo kan wan zhe bu xi liao le wo fei chang xi huan ta ban yan de jiao se</mandarin>
- Use the pinyin generator (https://pinyingenerator.dannyho1988.repl.run/) to generate the pinyin transcription in NSC’s format
- For words that have more than one pronunciation in Chinese, transcribe them as how they were pronounced even if consider ungrammatical in Standard Mandarin.
    - E.g. If “角色” is pronounced as “jiao3 se4” as opposed to the standard “jue2 se4” by the speaker, it should be transcribed as “jiao se” accordingly.
    - E.g. The construct “…了了 liao3 le4” in “我看完这部戏了了” is common in Singapore Mandarin and should be transcribed as “liao le” and not “le le”.

2. Transcribing Singapore Chinese expressions

- Singapore Mandarin is a variety of Mandarin spoken in Singapore that contains loanwords from Chinese dialects (Hokkien, Cantonese) English and other local languages such as Malay.
- Expression originating from Chinese dialects and other local languages are to be written in Roman alphabets, whereas local Mandarin expressions (i.e. 巴刹、巴士、巴仙、组屋、甘榜) are to be written in Chinese characters.
- More info on Singapore Chinese: https://zh.wikipedia.org/wiki/新加坡华语
    - Example sentence:
    - !walao! [eh] <mandarin>这个国家:zhe ge guo jia </mandarin> #sibei# #jialat# [leh] G_S_T <mandarin>明年要涨到九巴仙:ming nian yao zhang dao jiu ba xian</mandarin> [leh] <mandarin>去巴刹买菜一定会更贵以后我:qu ba sha mai cai yi ding hui geng gui yi hou wo</mandarin> #bo-lui# <mandarin>了啦:le la</mandarin>

3. Transcribing discourse particles in Singapore Chinese

- Singapore Chinese contains discourse particles from Standard Chinese and borrowings from other Chinese dialects (i.e. Hokkien or Cantonese) in Singapore. Particles considered non-standard Chinese (usually borrowings from Hokkien or Cantonese) are to be written with square brackets, while those considered Standard Chinese are not to be indicated with any notation symbols.
- Examples of discourse particles used in Standard Chinese and Singapore Chinese.

Discourse particles in Standard Chinese (标准中文语气助词)	Discourse particles in Singapore Chinese (新加坡式华语语气助词)
啊、哦、呀、吗、嘛、啦、唉、呢、吧、哇、呀	[leh], [meh], [lor], [har], [hor]

* List is not exhaustive

Example sentence:

and I tell you **[ah]** <mandarin>这个人真的很不讲道理:zhe ge ren zhen de hen bu jiang dao li</mandarin> **[leh]** <mandarin>不对:bu dui</mandarin> **[meh]** <mandarin>难道我看错人了吗:nan dao wo kan cuo ren le ma</mandarin> do you think am I correct or not [ha]

4. Representing Singapore Mandarin utterances in transcription

- For incompletely uttered word in Chinese, type the complete character that bests represent the speaker's intended utterance.
- Spoken dialogue: "其实 xia~… 其实我觉得小明人真好"

Correct	其实小其实我觉得小明人真好
Incorrect	其实 xia~ 其实我觉得小明人真好 (type in pinyin)
Incorrect	其实消其实我觉得小明人真好 (sentence is incorrect as 消 is inappropriate to the context of the dialogue even though it sounds similar to 小)

- For incorrectly pronounced words, type the actual character intended by the speaker if pronounced slightly incorrect (e.g. tone is off or due to accent)

E.g. 1 Spoken dialogue: "王乙康部长 (zhang1)声明…"

Correct	王乙康部长声明
Incorrect	王乙康部樟声明

E.g. 2 Spoken dialogue: “蓝教练是男教练…（nan2 jiao4 nian4 si4 lan2 jiao4 nian4）”

Correct	蓝教练是男教练
Incorrect	南教念寺蓝教念

- For incorrectly pronounced words, type the approximate/ intended characters by the speaker if the pronounciation sounded nothing like the original expression.

Spoken dialogue: “王乙凯部长声明…” (the speaker mispronounced 康 as 凯)

Correct	王乙凯部长声明
Incorrect	王乙康部长声明

- For non-existent word (due to verbal error), type the approximate / intended characters appropriate to the context of the dialogue

Spoken dialogue: “我不懂他们要怎么 an1 tuo3， I mean 好好 arrange 那个东西啦…”

Correct	我不懂他们要怎安妥 I mean 好好 arrange 那个东西啦
Incorrect	我不懂他们要怎俺托 I mean 好好 arrange 那个东西啦

- Chinese discourse particle should be used if the preceeding utterance is in Chinese. Roman alphabets should be used if the preceeding utterance is in English.

Correct	我跟你讲啊， he’s very bad one [leh]，你觉得我这样说有错吗。
Incorrect	我跟你讲 [ah]，he’s very bad one 咧，你觉得我这样说有错 [mah]。

- For single English codeswitched word embedded in a Chinese discourse, the following discourse particle (if Standard Chinese) should be written in Chinese character.

Correct	你觉得我这样讲 correct 吗
Incorrect	你觉得我这样讲 correct [mah]

Annex E: Malay Style Guide

As a rule, please transcribe Malay in proper Bahasa Melayu, except in these cases:

1. Conjoined/ shortened Malay words

- Spell as spoken (for most).
    - Rule of thumb: if the shortened words sound weird (or too formal) when you spell it out, keep the shortened form. (i.e. No one says “begitu/begini” in colloquial Malay speech unless on purpose.)
    - ***Examples:***
        - dah, nak, tak, takde, takpe, tu, ni, gitu, gini, abeh, amacam, pastu, kat

2. Malay discourse particles

- Malay discourse particles should be tagged separately and not be included in the <malay> tag as they will go undetected
- ***Examples:***
    - [jer], [kan], [pe], [tau]

Correct	<malay>macam tahu je</malay>
Incorrect	<malay>macam tahu</malay> [jer]

3. Spacing

- Do not leave a spacing after opening and/or before closing of the <malay> tag.

Correct	<malay>semalam kau pergi mana</malay>
Incorrect	<malay> semalam kau pergi mana </malay> <malay> semalam kau pergi mana</malay> <malay>semalam kau pergi mana </malay>

4. Names/ titles

- Malay names/ titles can be included in the <malay> tag if they are mentioned in a string of Malay sentence

Correct	<malay>aku rasa haji hussin kat sana</malay>
Incorrect	<malay>aku rasa</malay> #haji hussin# <malay>kat sana</malay>

- But if they are in between an English sentence, they should be hashtagged

Correct	I think #datin rosmah# is there
Incorrect	I think <malay>datin rosmah</malay> is there

5. Hybrid words

- When English rules are attached to Malay words (e.g. “barang-s”) or Malay rules are attached to English words (e.g. “ber-happy”), they should be hashtagged
- Note: hashtag the entire term and do not use hyphens

Correct	<malay>mari kita</malay> #berhappy#
Incorrect	<malay>mari kita</malay> #ber-happy# <malay>mari kita</malay> #ber# happy <malay>mari kita</malay> #ber-# happy

Annex F: Tamil Style Guide

Tamil Style Guide

1. Conjoined/ shortened Tamil words:

- Spell as spoken (for most).
    - Rule of thumb: if the shortened words sound weird when you spell it out, replace it with proper tamil

Correct	அப்புறம்
Incorrect	அப்பிரம்

2. Spelling errors

- Generally take note of spelling errors for the words spoken by the speakers because they will sometimes use proper written tamil words as it is or they will use the colloquial spoken words in their sentences, so, please check if the spelling is correct in either term.

Correct	-ஊரு -அவற்றை -தண்ணீரினால்/தண்ணீர்னால -சொல்லுறீங்க -பண்ணுறீங்க அதுக்கு/அதற்கு
Incorrect	-உரு -ஆவற்றை -தண்ணீர்நால -சொல்ரீங்க -பன்ரீங்க -அதர்க்கு

3. Tamil discourse particles

- Tamil discourse particles should be tagged separately and not be included in the <tamil> tag as they will go undetected:
- ***Examples:*** [dah], [dey], !aiyoyo!

Correct	!aiyoyo! <tamil>அப்புறம் என்ன ஆச்சு:appuram enna aachu</tamil>
Incorrect	<tamil>ஐயய்யோ அப்புறம் என்ன ஆச்சு:aiyaiyo appuram enna aachu</tamil>

4. Spacing

- Do not leave a spacing after opening and/or before closing of the <tamil> tag.

Correct	<tamil>என்ன கொடுமை:enna kodumai<tamil>
Incorrect	<tamil> என்ன கொடுமை:enna kodumai <tamil>

5.Use new format for tamil converter: https://tamilconverter.dannyho1988.repl.run/

Correct	<tamil>அவன் அவனோட:avan avanooda</tamil>
Incorrect	<tamil>அவன் அவனோட:avn avanoda</tamil>

6. Names/ titles

- Tamil names/titles should be included in the <tamil> tag if they are mentioned in a string of tamil sentence

Correct	<tamil>நீ கவிதா கிட்ட சொல்லிட்டியா:nii kavithaa kitta sollitiyaa</tamil>
Incorrect	<tamil>நீ:nii</tamil> #kavitha# <tamil>கிட்ட சொல்லிட்டியா:kitta sollitiyaa</tamil>

- But if they are in between an English sentence, they should be hashtagged

Correct	[dey] is #kavitha# here
Incorrect	[dey] is <tamil>கவிதா</tamil> here

7. Hybrid words

- When English rules are attached to Tamil words (e.g. “saappiduraa-ning”) or Tamil rules are attached to English words (e.g. “egg tart-aa), they should be hashtagged

Correct	<tamil>நீ என்ன சாப்பிடுற:nii enna saappidura</tamil> #egg tart-aa#
Incorrect	<tamil>நீ என்ன சாப்பிடுற:nii enna saappidura</tamil> egg tart [ah]

8.Careless mistakes that will create error in sentence/ change the meaning of sentence

- When certain words by the speaker are mispronounced replace the correct word that fits the context. Also check for spelling errors to avoid change in meaning of a sentence

Correct	-சரி என்ன வேலை வேண்டும் என்றாலும் -வேலை போனதிலிருந்து

	-அதை தவிர வந்து -திருப்பி யாரயாவதை -ஒரு பெண்/ ஒரு பொண்ணு
Incorrect	-சரி என்ன வேலை வென்றும் என்னாலும் -வேலை பொன்னத்தில் இருந்து -அதற்கு தவிர வந்து -திருப்பி யாரவதை -ஒரு பொன்னு