Data release for 2019 Cognition article, coding principles for:

Alderete, John and Queenie Chan 2018. Simon Fraser University Speech Error Database Cantonese 1.0 (SFUSED Cantonese 1.0). Data collection, Simon Fraser University.

Goal: make clear all conventions, stylistic or otherwise, used in classifying speech errors that have already been imported. See the dataInput document for conventions for error submissions.

Fields: in the spreadsheet that constitutes this data release, the header gives the field names. Fields are of various types, e.g., Example, Class, Word, which characterize the nature of the variable, and are the first part of the field name. For example, classMasterType is the basic characterization of the error type, and classContextual is a class field that distinguishes between contextual (Y) and noncontextual errors (N). Generally the field names in the headers are self-exlanatory, but when they are not, see the comment for further explanation.

Words vs. phrases (Packard 2016) A Cantonese word is a minimal free form that can stand alone. A Cantonese phrase is a construction made of two or more words that can independently occupy different part-of-speech slots.

Monosyllabic morphemes The following table illustrates four types of word component monosyllabic morphemes in Cantonese, of which only two types can be free (i.e. stand alone).

	Contentful/ functional?	Free/ bound?	Examples
Free morphemes	Contentful	Free	Noun: 水 sœi35 ‘water’, 車 tse55 ‘car’
			Verb: 食 sik22 ‘eat’, 走 dzau35 ‘leave’
			Adj.: 紅 hoŋ21 ‘red’, 快 fai33 ‘fast’
Bound roots	Contentful	Bound	Noun: 巾 gan55 ‘towel’ (cf. 毛巾 mou21gan55) Verb: 鍾 dzoŋ55 ‘like’ (cf. 鍾意 dzoŋ55ji33) Adj.: 彩 tsoi35 ‘lucky’ (cf. 好彩 hou35tsoi35)
Functional morphemes	Functional	Free	Sentence final particles? (Not sure. We need to talk about this.)
Affixes (cf. Synopsis §3.2, §7)	Functional	Bound	Prefix: 唔- m21 ‘negative’ Infix: -鬼- gwai35 ‘for emphasis’ Suffix: -哋 dei22 ‘plurality’

Multisyllabic words

Composition (Pac 72) is a word formation process that combines two bound roots, or combines a bound root with a free morpheme. A word formed via composition must contain at least one bound root (cf. bisyllabic words under “Bound roots” in the table above).

Compounding (CRG 58-62) is a word formation process that combines two (or more) free words. The process creates compound nouns/verbs/adjectives. Verb-object compounds (compound verbs) can sometimes be confused with phrases consisting of a verb and a direct object (verb phrases), since in both constructions, aspect markers and verbal particles are allowed to intervene between the verb and the noun. Below are some properties that differentiate a verb-object compound from a verb phrase:

(i) the resulting meaning of a verb-object compound often corresponds to an intransitive verb in English and does not fully reflect that of the object (e.g. 皺眉頭 dzau33mei21tau21 ‘lit. wrinke-eyebrow = frown’).

(ii) a verb-object compound has an idiomatic meaning (e.g. 飲茶 jam35tsa21 ‘lit. drink-tea = have dim sum’).

Words and phrases in Intended, Error and Source Phonetic boxes

Spaces in these boxes are used to separate two words, so there is no space between the syllables of a compound.

Example: jyn21tsyn21(完全 ‘completely’ (compound)) vs. bei34 lei23 (俾你 ‘give you’ (phrase))

Notes:

a. Note the space between *bei34* and *lei23.*

b. If a source sound is inside the Error word (syllables not separated by spaces), then the error is word-bounded, and the *Word-bounded?* field should be checked. Otherwise, the error is not word-bounded.

Words and phrases in Intended, Error and Source Orthographic boxes

Chinese characters should occupy these boxes. Phonetic transcription should only be used when (i) a syllable is clipped, or when (ii) the syllable does not have a corresponding Chinese character.

No space (i) between Chinese characters, (ii) between a Chinese character and the phonetic transcription of a syllable, (iii) between phonetic transriptions of two syllables in these boxes.

The Example box (i.e., the longform of the error)

In Error Submissions made by researchers, the Example field in a spreadsheet resembles everyday Chinese writing: long strings of Chinese characters, not separated by any space, punctuated with commas ‘,’ and periods ‘.’ (Chinese periods ‘。’ are not used in error submissions). This section systematizes what appears inside the Example box in the database, which looks different from the Example field in error submissions.

Error Introduce an error with a forward slash “/” that is preceded by a space. If the error is a monosyllabic word, insert a space after (i.e., “_/Syllable_”). If the error is a mutiple-syllable word, insert a number symbol “#”, followed by a space (i.e., “_/SyllableSyllable#_”).

Example: 我想 /煎 讚吓呢個讀者先嘅. (intended: 先 sin55; error: 煎 dzin55)

Example: 如果 /艷格# 嚟講呀吓. (intended: 嚴格 jim21ga:k33;
error: 艷格 jim22ga:k33)

If the error word is monosyllabic and clipped, it should be suffixed with “=”, followed by a space. If the error is a multiple-syllable word and the final syllable is clipped, it should be suffixed with “=”, followed by a space. No number sign “#” should be inserted.

Example: 反而你 /dy= ^元朗 呀再掃落去後面呢.

(intended: 元朗 jyn21loŋ35; error: dy=; source: 元朗 jyn21loŋ35)

If the error word is followed immediately by a natural short pause, as indicated by a comma “,”, there should be no space between the monosyllabic error word and the comma, or between the “#” and the comma if the error has multiple syllables.

Example: /係 tsə22 咪#,掉番轉頭因為佢係咁個人.

(intended: 係咪 hai22mai22; error: 係 tsə22 咪; no source)

If the error word is preceded immediately by a natural short pause, as indicated by a comma “,”, there should be no space between the comma and the forward slash that introduces the error word, i.e., “xxx,/Error xxx”, but *“xxx,_/Error xxx”.

If the error word is at the end of a clause, as indicated by a period ‘.’, there should be no space between the monosyllabic error word and the period, or between the “#” and the comma if the error has multiple syllables.

Example: ...等佢 /b|wiŋ23 不超生#.

(intended: 永不超生 wiŋ23bat55tsiu55saŋ55;
error: b|wiŋ23 不超生 b|wiŋ23bat55tsiu55saŋ55; no source)

Source A source prefix “^” is prefixed to a word (not just the syllable) that contains the source sound. There should be no space before the source prefix “^”. There should be a space after the word containing the source sound (i.e., “PreviousWord^Source_NextWord”).

Example: 即係可以譬如 /然輕# 一輩^可以 買得起嘅.

(intended: 年輕 nin21hiŋ55; error: 然輕 jin21hiŋ55; source: 可以 ho35ji22)

Note: “^”is prefixed to the word 可以, not just to the second syllable 以 of the word, even though it is the second syllable that contains the source sound [j].

If the Source is a phrase made up of two or more words, there should be a space after the Source phrase, but no space within the Source phrase.

Example: 有小小自己俾個氹 /_踩= ts= 即係自己俾個氹 ^自己踩 咁樣.

(intended: 自己踩 ji22gei23tsai35; error: 踩 tsai35; source: 自己踩 ji22gei23tsai35)

If the source is within the error word, the source prefix “^” should be prefixed to the forward slash “/” that introduces the error. There should be space before the “^” and no space between the “^” and “/” (i.e., “PreviousWord_^/Error_NextWord”).

Example: 即係 ^/啲啲# 小趣聞嚟嘅.

(intended: 呢啲 li55di55; error: 啲啲 di55di55; source: 啲啲 di55di55)

If the source is immediately followed by a natural short pause, as indicated by a comma ‘,’, or is at the end of a clause, as indicated by a period ‘.’, there should be no space between the source word and the punctuations “,” or “.”.

Example: 我想問吓^呢, /gə55ai22# 又係問感情^啦.

Sources that are within a ten-syllable contextual window from the error should be identified.

Characters or transcription Chinese characters should always be used in the Example box, unless there are no Chinese characters corresponding to the syllables/words uttered. In cases where transcription must be used, the transcription should occur freely, not within square brackets.

Example: 突然之間, /jip35 會唔會係個凳底嗰度有啲問題.

(intended: 咦 ji35; error: jip35 (no corresponding Chinese character); no source)

Speakers If two speakers take turns in an Example box, their utterances should be introduced by the letters “A” and “B”, followed by a colon “:”. Speaker tags should not be used.

Example:

A: 同埋好^defensive 囉. B: 愉,愉景灣都係咁樣㗎喎類似. A: /愉頂灣# 就 I= 愉景灣好小小.

Each utterance should always end with a period “.” if it is a complete clause, with “…” if it is not a complete clause, or with “?” if it is a question.

... Start an example with ‘…’ if it is not the beginning of a clause. End an example with ‘…’ if it is not the end of a clause. No space between “…” and what precedes/follows.

If a speaker identifies herself and/or other speakers in the podcast by name, such names should be blocked out in the Example box by replacing them with “(talker name)”. If the error is contained within the name, leave it and check ‘Y’ for Personal Info? in bottom left.

Example: …我"(talker name)"繼續喺度要問吓啊初^呢 xxx /lwoŋ21 曉初# 講番我哋今日嘅戲目…

(intended: 黃曉初 woŋ21hiu35tso55; error: lwoŋ21 曉初 lwoŋ21hiu35tso55; source: 呢 le55)

N.b.: the longform of the error has been anonymized for talker names, replaced by XNameX in eight instances.

Colon ":" introduces what a speaker's utterance in an Example box that has multiple speakers. A speaker should always be represented by an English letter, which always precedes ":". No space between a letter and ":". There should be a space between ":" and an utterance, i.e. "A:_utterance".
xxx indicates long unnatural hestitation or stops. There should be space before and after "xxx", i.e. "_xxx_".
= indicates a clipped syllable. No space between a transciption and "=". Space after "=", unless it is followed by a "," or ".".

If the clipped syllable(s) is the first part of a word, indicate the missing syllable(s) with _, and suffix the error with "="

Example: 有小小自己俾個氹 /_踩= ts= 即係自己俾個氹 ^自己踩 咁樣.

(intended: 自己踩 ji22gei23tsai35; error: 踩 tsai35; source: 自己踩 ji22gei23tsai35)
Comma indicates natural pauses. No space before nor after ",". Commas should be typed with an English input method ("Canadian English"), not a Chinese input method ("Pinyin-Traditional").
No insertion of comma within an error word. Note observation of the pause in comment box.
Period "." indicates the end of a clause. An English period "." (as opposed to a Chinese period "。") should be used.
Question mark "?" indicates the end of a question. Question marks should be typed with an English input method ("Canadian English"), not a Chinese input method ("Pinyin-Traditional").
Space Generally no space between Chinese characters, except for the scenarios described in "Error", "Source" and "xxx" in this section. In cases where there is a space between syllables in the Error Phonetic box (see previous section), there should be NO SPACE between the corresponding Chinese characters in the Example box.
No space between a Chinese character and an English word. There should be space between two English words. No space between a Chinese character and a transcription. There should be space between transcription of two syllables in succession.

In summary, concerning where to insert space, except for the scenarios described in "Error" and "Source":

				Space?
Between	Chinese character	and	Chinese character	✗
	Chinese character		English word	✗
	Chinese character		transcription	✗
	English word		English word	✓

	English word		transcription	✓
	transcription		transcription	✓
	Error#		,	✗
	Error#		.	✗
	^SourceWord		,	✗
	^SourceWord		.	✗
	^		/Error	✗

		Space?
Before	xxx	✓
	,	✗
	.	✗
	?	✗
	=	✗
	…	✗
	:	✗
After	xxx	✓
	,	✗
	.	✓
	?	✓
	=	✓
	…	✗
	:	✓

Sources in Complex Set of Processes
Assumption 1: The sources in the Source field should be consistent with the specific type of structure specified in the Intended and Intruder sound fields. Concretely, if there is consonant substitution in the sound fields for Intended and Intruder, then the Source will be specifically for consonants. Sometimes, there are two different structures, e.g. tone and consonant, and there isn't a single source for both structures. Basically we are looking for consistency in the sound fields, and the Source box in the Example box is consistent with the sound fields.
Assumption 2: In the larger Example box, all sources, even if they are for different structures, get a '^' prefix. If there is Source in the Example box that is not recognized in the sound fields, we state it as an assumption in the Comment box.

Sources containing multiple source sounds
Assumption: In the case where a Source word contains two (or more) syllables that contain the intruder sound, we take the syllable that matches the position of the Error.

Funny/non-native tones

In the case where a data collector thinks a tone is non-native (not one of the six basic tones in Hong Kong Cantonese), re-listen to the audio, and decide if:

1. the tone could actually be a native tone that is slightly off but still within the normal range. If so, the tone would not be an error and the entry should not be kept.
2. the non-native tone could be due to co-articulation, e.g. perseveration or anticipation of another neighbouring tone. If so, the tone would not be an error and the entry should not be kept.
3. if the none is not a within-range native tone (see step 1 above) or a tone influenced by co-articulation (see step 2 above), then the tone could be a phonetic error. In this case, the Master Type should be phonetic error, Type should be gradient, Phonologically illegal? should be checked 'N', and Phonotactic Violation should be left blank.

References

Packard, Jerome. Lexical word formation. In Chu-Ren Huang and Dingxu Shi (eds.), *A Reference Grammar of Chinese*, 67-80. Cambridge: Cambridge University Press.