Published in final edited form as:
*Behav Res Methods*. 2019 June ; 51(3): 1131–1144. doi:10.3758/s13428-018-1043-6.

# Cantonese AphasiaBank: An annotated database of spoken discourse and co-verbal gestures by healthy and language-impaired native Cantonese speakers

**Anthony Pak-Hin Kong**[1,*] and **Sam-Po Law**[2]

[1]Department of Communication Sciences and Disorders, University of Central Florida, Orlando, FL, USA

[2]Division of Speech and Hearing Sciences, University of Hong Kong, Hong Kong SAR

## Abstract

This paper reports the construction of a multi-modal annotated database of spoken discourse and co-verbal gestures by native healthy speakers of Cantonese and individuals with language impairment, Cantonese AphasiaBank. This corpus was established as a foundation for aphasiologists and clinicians to design and conduct research investigations into theoretical and clinical issues related to acquired language disorders in Chinese. Details in terms of the purpose, structure, and levels of annotation of the database (containing part-of-speech annotated orthographic transcripts with Romanization and corresponding videos) were described. The discussion presents the challenges of building a spoken database of a language that is not linguistically well-researched and without a proper written form for many of its lexical items, as well as how these issues are addressed. More importantly, the paper highlights the potential of Cantonese AphasiaBank as a powerful research tool to linguists and psycholinguists.

### Keywords

language database; Cantonese; Chinese; aphasia; discourse; gestures; stroke

## Introduction

The value of language databases to linguistic and especially psycholinguistic studies has been well-recognized and demonstrated by researchers in these fields. This paper introduces a newly developed spoken language database, the Cantonese AphasiaBank (Kong & Law, 2010–2014), that has been made available in the public domain since 2016. The database, containing discourse samples from unimpaired native Cantonese speakers and right-handed people with aphasia (PWA) living in Hong Kong of different ages and education levels, is first of its kind in Asian languages for studying spoken narratives. The PWA in the database include different aphasia types caused by stroke. Significantly, it encompasses not only

[*]Corresponding author: Anthony Pak-Hin Kong, Ph.D., Department of Communication Sciences and Disorders, University of Central Florida, Room 106, Health and Public Affairs II, Orlando, FL 32816-2215, antkong@ucf.edu, Tel: (407) 823 4791 Fax: (407) 823 4816.

information on distinctive linguistic properties in discourse production, but also non-verbal behaviors (i.e. co-verbal gestures) produced by healthy speakers and PWA. The elicitation protocol follows the English AphasiaBank protocol (MacWhinney, Fromm, Forbes, & Holland, 2011), but with careful adaptation to the local Chinese culture. This corpus was established with the premise that it would provide the necessary foundation for aphasiologists and clinicians to design and conduct research investigations of theoretical and clinical issues related to acquired language disorders in Chinese. The overarching goal of constructing this database is to improve the planning of assessment and remediation procedures for Chinese-speaking PWA worldwide, including those living in North America, through actively sharing this multimedia database with any clinicians and researchers who work with Chinese speakers with acquired language deficits. However, as Cantonese AphasiaBank contains both normal and disordered language data, we will demonstrate that it constitutes a rich resource for addressing various issues of language behaviors that are of interest to linguists, neurolinguists, and psycholinguists.

Generally speaking, corpora drawn from written texts outnumber those based on the spoken language. This is also the case for Chinese as revealed in a survey of Chinese language corpora for research (Yang, 2006), in which about 70% of the listed corpora were constructed using a written source. The Cantonese AphasiaBank was designed to distinguish from these databases in several ways. Not only was its scale larger than most corpora in terms of the number of speakers, but also its scope of content that included both unimpaired language and aphasic discourse production. In addition, information on verbal language performance ranging from the lexical, clausal/sentential, to discoursal levels were systematically captured and presented, along with performance of co-verbal gestures that are synchronized with discourse language production. Finally, unlike most previous databases that were based on open-ended conversations, the participants in this database performed an identical set of discourse tasks with target contents controlled. As a result of these unique features, the Cantonese AphasiaBank allowed investigations by linguists, neurolinguists, and psycholinguists. Apart from the above, the database has also supported education of student-clinicians of speech-language pathologists[1].

In order to better appreciate how the Cantonese AphasiaBank is advantageous for research applications and, therefore, a new contribution to current resources, a brief introduction of stroke-induced aphasia and a review of existing spoken Chinese databases is in order. Stroke has been listed as one of the target diseases on the global agenda for prevention and control by the World Health Organization. The burden of stroke is particularly serious in Asian countries (Kim, 2014) because of their rapidly aging population. The prevalence of post-stroke aphasia in the Indo-European populations is about 40% (Salter, Teasell, Foley, & Allen, 2014; Wade, Hewer, David, & Menderby, 1986). Applying this prevalence rate to Chinese speakers (due to the lack of comparable figures available for Asian countries), one can see a huge demand for language rehabilitation from these individuals. However,

[1] Since 2014, academic and clinical staff at the University of Central Florida, The University of Hong Kong, Polytechnic University of Hong Kong, The Education University of Hong Kong, and Hong Kong Institute of Vocational Education have utilized the database as part of their instructional materials for topics related to acquired language disorders. Extension to clinical training for practicing speech-language pathologists and related medical health-care professionals is also suitable. It is believed that research and clinical applications of this shared database will grow with an increasing number of users.

compared to the rigorous research agenda of investigating and evidence-based protocols for managing aphasia in English, there is a great paucity of resources reported for use in Chinese-speaking PWA (Kong, 2017). Chinese PWA is one of the underrepresented minority groups as listed by National Institutes of Health. Traditionally, aphasiology research has employed methods of single case studies, case series, or participants in groups of small sizes (Martin & Kalinyak-Fliszar, 2014; Willmes, 2007). Using big data to predict epidemics, address pathological deficits, and improve quality of life has become an upcoming trend in the healthcare industry, including the field of speech therapy or communication sciences and disorders (Faroqi-Shah, 2016). Many clinical as well as research questions can only be answered with data from substantial numbers of patients, their performances across different language tasks, and responses to individual test items. The Cantonese AphasiaBank project (Kong & Law, 2010–2014) was initiated to go beyond the conventional narrow-sampling approach of investigation.

The Chinese language family consists of seven major dialects including Mandarin, Min, Hakka, Wu, Cantonese, Xiang, and Gan. According to a number of reviews of spoken Chinese databases (Chui & Lai, 2008; Leung & Law, 2001; Wang, 2001; Yang 2006), there are currently a total of 13 in Mandarin, and ten in spoken Cantonese. Noted that Cantonese is the second most widely spoken dialect with over 52 million speakers distributed over southern China and overseas Chinese communities; it also differs from Mandarin in that it has both the spoken and corresponding written form. Similar to corpora in other languages, Chinese spoken databases serve different purposes and may represent different registers or genres from different sources. Some consist of recordings of read single words, utterances and passages from printed materials in Mandarin (e.g., Chou & Tseng, 1999) or in Cantonese (e.g., Lee, Lo, Ching, & Meng, 2002) for developing speech recognition and synthesis technology; some contain pre-planned dialogues from television dramas in Mandarin (e.g., Lee, 2011) or in Cantonese (e.g., Xu & Lee, 1998), while others contain speech of a more spontaneous nature from telephone conversations in Mandarin (e.g., Zhou, Li, Yin, & Zong, 2010), Cantonese radio programs (e.g., Leung & Law, 2001), and monologues such as storytelling in Mandarin (e.g., Chafe, 1980) or in Cantonese (e.g., Chui & Lai, 2008). The above-mentioned databases were constructed solely with language samples of unimpaired speakers. Moreover, videos files capturing speakers' performance during time of speech sample collection were absent since the majority of language materials in these databases were based on audio recordings[2]. Furthermore, only two of these Cantonese adult corpora, namely the Hong Kong Cantonese Adult Language Corpus (HKCAC; Leung & Law, 2001; Leung, Law, & Fung, 2004) and the Hong Kong University Cantonese Corpus (HKUCC; Luke & Nancarrow, 1997; Wong, 2006a) have part-of-speech (POS) tagging[3].

In the rest of this paper, we describe the background of building Cantonese AphasiaBank. Specific search functions of this database, its multi-modal display features, as well as

[2] It is not hard to understand why there is an imbalance between written and spoken databases. Printed materials are readily available and can be easily processed for linguistic annotation. In contrast, spoken databases require acceptable to high quality of recordings, followed by a labor intensive stage of orthographic and/or phonetic transcription before further work can be done.

[3] Whereas most of the corpora are of adult speech, some are child language databases for language acquisition studies (e.g., Tardif, 1993; Lee & Wong, 1998).

specific challenges of and solutions to annotation of database contents will also be illustrated. We will also demonstrate how Cantonese AphasiaBank can be used by researchers of different language disciplines as a research tool that can subsequently facilitate the management of Chinese-speaking PWA.

## Cantonese AphasiaBank

There are two corpora in Cantonese AphasiaBank, CANtonese corpus of Oral Narratives (CANON) and Database of Speech and GEsture (DoSaGE). They can be accessed at http://www.speech.hku.hk/caphbank/search/. After registration, users can utilize all database contents they wish to analyze (see Figure 1 for the ‘About us’ and ‘Log in’ screen).

### CANtonese corpus of Oral Narratives (CANON)

CANON differs importantly from HKUCC in a number of aspects. While HKUCC contains mainly conversations of a variety of topics among highly educated young and middle-aged Cantonese speakers and HKCAC contains conversations recorded from phone-in programs on the radio, data in Cantonese AphasiaBank are monologues collected from local native speakers of Cantonese residing in Hong Kong (a linguistically homogenous city of China) balanced in gender, age, and education. To be specific, the first corpus, namely CANtonese corpus of Oral Narratives (CANON), contains annotated orthographic and morphological information as well as romanized transcripts of 149 unimpaired speakers and 105 PWA. The language elicitation protocol included (1) description of a single color photo displaying the scene of rescuing someone in a flood, (2) description of a single black and white line drawing of “Cat Rescue”, (3) two sequential picture description tasks using two sets of black and white drawings – “Broken Window” and “Refused Umbrella”, (4) a procedural discourse task of describing how to prepare an “Egg and Ham Sandwich”, (5) telling of two stories – “The Boy Who Cried Wolf” and “Tortoise and Hare”, and (6) a personal monologue of an important event. For all PWA, the protocol included (7) an additional monologue task of telling their “stroke story”. Only tasks (1) to (5) were elicited using pictorial materials. If needed, task-specific probing questions were also given. Apart from the above narrative tasks, each PWA is administered language tests to assess repetition of words and phrases, noun and verb naming, and (non-)verbal semantic skills as well as the Action Research Arm Test (ARAT; Lyle, 1981) to quantify the degree of upper limb hemiplegia. The demographic information of both speaker groups is given in Tables 1 and 2. Figure 2 shows the interface displayed to users for selecting subject using the embedded filter features in this database (including “subject type”, “aphasia type”, “gender’, “age”, “education level”, and “ARAT score”).

As one can see, the data of CANON are not free language samples, that is, the target content of spoken output was controlled to be the same across our participants. The advantage of task specific data is to enable us to have control over the content such that one may study how language output differs as a function of age, gender, and education. The language samples were orthographically transcribed by two linguistically trained research assistants. Inter-rater reliability of orthographic transcription was computed by randomly selecting 10%

of the samples and double-checked against the audio recordings by the project investigators, and was found to have an agreement of greater than 99%.

### Database of Speech and GEsture (DoSaGE)

The second corpus in Cantonese AphasiaBank is the Database of Speech and GEsture (DoSaGE) that contains digitized video recordings of procedural discourse, story-telling, and personal monologue from 131 (of the 149) unimpaired speakers and 96 (of the 105) PWA in CANON. These videos are synchronized with the corresponding orthographic transcripts, using EUDICO Linguistic Annotator (Lausberg & Sloetjes, 2009), with independent annotation of the forms and functions of all co-verbal gestures (Table 3).

## Markup, search, and annotation in CANON

This section describes the characteristics, structure, and annotation of the data in CANON. Note that the morphological tagging of parts-of-speech (POS) in CANON is automatic, unlike the manual annotation in HKUCC, with specific tagging rules written for CANON to reflect morphological processes of Cantonese.

For each participant in Cantonese AphasiaBank, the orthographic transcriptions of all language samples are combined into one single document with a code name. The identities, personal and demographic information of the participants, and times and dates of individual recordings are kept in a separate master file. Each narrative in the transcript begins with a line marking the name of the sample. Transcriptions are formatted using the Codes for the Human Analysis of Transcripts[4] (CHAT; MacWhinney 2000). The transcripts in CANON are linked to audio and video recordings through a computerized analytic program named Child Language Analyses (CLAN; MacWhinney, 2003). CLAN allows one to carry out a variety of linguistic analyses, such as frequency count, lexical diversity, mean length of utterance (MLU), as well as searches for user-specified combinations of words, character strings, or words in context, etc.

In addition to orthographic transcription, each lexical entry or token demarcated by a space is annotated for POS, an automatically generated phonetic transcription in Cantonese romanization, and an English gloss. For each transcript, segmentation of utterances largely follows the "one verb or clause per line" principle, except when a verb subcategorizes for a clause. Specifically, a new utterance (therefore a new line in CLAN) is formed (1) when a speaker restarts, partially repeats what s/he just said, (2) when a speaker switches to a new topic or when the topic contains more than a noun phrase (i.e., a clause), (3) when there is an interjection, a connective, or a filler between clauses, or (4) whenever the adverbial 跟住(呢) 'then' is used. Examples of two utterances are given in Figure 3, where (a) shows morphological annotation with unambiguous tagging, (b) illustrates initial POS tagging with some tokens having multiple tags separated by the symbol "^" (shown in shaded text), and POS annotation after the transcript has been manually verified. In each example, the first tier

[4] The CHAT format was first developed during 1984–1988 by a group of developmental psycholinguists led by B. MacWhinney. The system was converted to a format compatible with the XML Internet data format between 2001 and 2004 in collaboration with the Linguistic Data Consortium (1992). CHAT has been adopted by two child language corpora in Cantonese (Fletcher, Leung, Stokes, & Weizman, 2000; Lee & Wong, 1998).

represents orthographic transcription, where tokens are divided by a space. The second tier consists of information on POS, Cantonese Romanization and an English gloss of each token, and in that order.

The morphological (POS) tagset in CANON has a total of 38 classes as listed in Table 4, fewer than the set of 54 tags in HKUCC, another adult Cantonese corpus mentioned earlier. While the two tagsets share many common classes as expected, the main differences lie in that (i) derivational and inflectional morphemes are distinguished in terms of their position of occurrence in HKUCC but classified under the category of “affix” in CANON; (ii) major content word classes, i.e., nouns, verbs, and adjectives, are distinguished in terms of number of constituent morphemes and furthermore in internal structure for verbs in CANON but not so in HKUCC; (iii) the classes of “short form”, “fixed expression” and “idioms” in HKUCC fall in the category of “expression” in CANON; and (iv) the classes of time, pronoun, numeral, verb, adjective and noun morphemes in HKUCC do not have counterparts in CANON. Although seven of the eight narratives from each participant are elicited using specific stimuli, the variety of lexical forms in the language samples is still evident. A total of 4,450 entries are listed in the CANON dictionary[5].

The current study applied the MOR tagger, a computational systems for the morphosyntactic analysis of the spoken language data in the CHILDES (http://childes.psy.cmu.edu) and TalkBank (http://talkbank.org) databases, for automatic annotation of POS. This procedure generally followed those listed in the English AphasiaBank project (MacWhinney & Fromm, 2016). According to MacWhinney (2012), the MOR grammars have been built for Indo-European languages, such as English, Spanish, French, Italian, Dutch, and German, as well as three Asian languages, including Mandarin, Cantonese, and Japanese. Specifically, the Cantonese MOR tagger (http://talkbank.org/morgrams/) was designed based on statistical distribution of lexical items in Cantonese and then with contextual rules of Cantonese grammar. While the MOR tagger reaches 98% accuracy on English adult corpora (MacWhinney, 2012), the tagger in the CANON has a comparable but slightly lower tagging accuracy (based on the total token of 132,024).

Users of Cantonese AphasiaBank who want to conduct a quick and simple search of specific lexical information of the database transcripts can do so through the “Search by Word” tab. The search can be performed by specifying one or more of the following parameters: a particular Chinese character, a specific Chinese lexicon (that is made up of one or more Chinese characters), jyutping (a romanization system for Cantonese developed by the Linguistic Society of Hong Kong), part-of-speech tags of transcription, glossary in English, and narrative task(s) that contain the search item. Figure 4 displays a screenshot of the interface displayed to users for defining search criteria on this tab. The length of utterances to be displayed in the results can also be pre-set by adjusting the concordance length parameter.

A search for a simple Chinese keyword 仔 ‘son’ is illustrated here. This character, apart from acting as a noun ‘son’ in oral narratives, can also serve multiple functions such as a suffix or

[5] Note that this dictionary is expected to increase further in size when additional data are collected and included in the future.

a constituent of a compound noun or proper noun. If a user is interested in knowing the total number of lexical items containing this character in Cantonese AphasiaBank, the frequency of occurrence in the database for each items and its part-of-speech, and/or how each lexical item was used by PWA (versus controls) across different narrative tasks, a simple “Search Keyword (Chinese)” result displayed in Figure 5 demonstrates the following: (a) there is a total of 1,303 tokens of lexical items that contains the character 仔, (b) there is a total of 69 different lexical types with different POS across the various narrative tasks within the database, and (c) the screenshot shows three of the 69 lexical types in the “Word” column – suffix in 几仔 ‘small chair’, compound noun constituent in 細路仔 ‘child’ and 男仔 ‘boy’, and single noun in 仔 ‘son’. Note that this display of frequency of occurrence is different from the usual notion of frequency count in most databases of open-ended content, which is supposed to reflect extent of word usage. The textbox displays a list of utterances in which the keyword 仔 appears. The tags (containing extra information pertaining to POS, jyutping, and gloss) can be removed by clicking the “Show/hide Tags” button. Each of these lines can also be clicked to further examine the language sample in which the target keyword is found. Note that there are four different matching modes, namely “Exact”, “Starts with”, “Contains”, and “Ends with” that can be selected in the “Match Mode” drop-down menu, for result filtering. In particular, the default mode of “Exact” search will only include a list of utterances where the exact same keyword appears in the database. For “Starts with”, any lexical items starting with the keyword entered will be identified. “Contains” mode will search for any lexical items that contain the keyword entered, and “Ends with” mode will identify any lexical items ending with the keyword entered. All users can then download their search results in the format of a Microsoft Excel worksheet. Importantly, these search results can be useful for linguists and/or psycholinguists who have specific research questions to address or hypotheses to test, and find our data collection methods and data presentation suitable for their purposes.

## Search and display of materials in DoSaGE

The video recordings of the subjects performing all discourse production tasks can be viewed in the “Browse Videos” tab, where a simple search form (drop-down menu of “Subject type”, “Aphasia type”, and “Task”) can be found. The video will be played on the left, while a moving window of transcriptions will be displayed on the right. The current sentences spoken by the subject in the video will be highlighted in yellow. For the tasks of procedural discourse, story-telling, and personal monologue, synchronized annotation information of gestures employed by the speaker is available and highlighted in orange (see Figure 6). In other words, the video, display of language transcriptions, and gesture annotations are time-framed (with information on duration of pauses) and synchronized. The detailed multi-level and multi-modal performance of one’s discourse production is, therefore, readily available for inspection. The transcribed texts can also be clicked for the video to skip to the exact position a user desires to view. Finally, a watermark with the user’s login email address, IP address, as well as time and date of viewing is automatically generated to avoid unauthorized duplication of our database videos.

## Specific challenges to building a Cantonese spoken corpus with POS annotation

There are many challenges in constructing a spoken corpus in Cantonese. We discuss below how we handle these difficulties and those that are unique to the language at every stage of the development of CANON, including challenges associated with transcribing and annotating spoken output of PWA.

### Orthographic representation of colloquial morphemes

Cantonese is essentially a spoken language; therefore, the problem of orthographically transcribing colloquial morphemes in the language is obvious. While written forms for these items can be found in the popular culture, such as local magazines and comic books, there is no standardization. To deal with this problem, we consult online Cantonese dictionaries including 粵語審音配詞字庫 (Chinese Character Database with Word Formations) at the Chinese University of Hong Kong (http://humanum.arts.cuhk.edu.hk/Lexis/lexi-can/) and CantoDict v1.4.2 (http://www.cantonese.sheik.co.uk/scripts/wordsearch.php?level=0). For morphemes not found in these sources, Cantonese Romanization would be used. As CLAN supports Unicode, colloquial characters not available in Unicode are represented either in Romanization or by an orthographically similar homophonous character; for instance, the progressive aspect marker 𡁵 which cannot be found in Unicode is written as 緊 in CANON.

### Orthographic transcription of homophonic and homographic morphemes, allomorphs, and fused syllables

Cantonese morphemes, similar to Mandarin, have a high degree of homophony, and many of which are written with the same characters, resulting in ambiguity in tagging. This is particularly problematic if the morphemes in question are highly frequent, e.g., 嘅 [ kE3] as a possessive marker, a relative clause marker, and a sentence final particle. Our interim solution is to distinguish the different homographic homophonous morphemes with a numeric code following the character, 嘅 1, 嘅 2, 嘅 3. Conversely, some Cantonese morphemes have alternative acceptable phonological forms, e.g. [ lIN1] and [ lIk1] for ‘carry’. In suchcases, different characters, 拎 [ lIN1] and 搦 [ lIk1], would be used to represent the various forms as much as possible.

A related issue is the occurrence of fused syllables, which is considered a unique feature of Chinese phonology (Wong, 2006b). Different extent of fusion from reduction to a single syllable to deletion of coda or onset can occur for different syllable strings. A single character would be used if available, e.g. 咩 [ mE15] in Figure 1(b) for 乜嘢 [ mAt1 jE5] ‘what’, where the rime of the first and onset of the second syllable are deleted.

### Line segmentation and tokenization

Partly due to the lack of word boundaries in Chinese text, the definition of a word in Chinese is far from straightforward. As characters represent morphosyllables, our guiding principle is to treat each character as a separate token, except for affixation and compounds which may appear in the categories of nouns, verbs, and adverbs. We consult the online dictionary of HKUCC for compounds in the language. After tokenization, the discourse is segmented into units for further analysis. While some spoken corpora would define an utterance in terms of prosodic boundaries (e.g., Zhou et al., 2010), our segmentation is syntactically based in anticipation for building a syntactic parser. A language sample is divided into separate lines in the transcript with each one corresponding to a clause as much as possible.

## POS of grammatical morphemes

One major difficulty in grammatically annotating a Cantonese corpus is POS classification, since less systematic and theoretical research has been conducted on the Cantonese grammar, compared with Mandarin. As known among Chinese linguists, determining the POS of grammatical morphemes is particularly challenging for the so-called verbal particles and affixes because many of them (e.g. prepositions or coverbs) originate from content words. We adhere to two principles when classifying candidates for the two classes. The content word status of a morpheme is maintained unless it is a bound morpheme, e.g., 士 [ si6] 'scholar' (somewhat equivalent to the English suffixes ' –ist' in 'psychologist', '–er' in 'philosopher') in 護士 'nurse', 師 [ si1] 'professional people' in 會計師 'accountant', and/or it is semantically empty or detached from its origin, e.g., 仔 [ tsAI2] ('son' when standalone) as an affix meaning 'diminutive' in 簿仔 'a little notebook', 頭 [ tHAU4] ('head' when standalone) 'suffix without meaning' in 石頭 'rock-suffix'. As such, 落 [ lOk6] 'down' in 行落去 'walk-**down**-go', 埋 [ maI4] 'approach' in 行埋嚟 'walk-**approach**-come' are treated as directional verbs, and 爛 [ lan6] 'broken' in 搣爛 'tear-broken' as the resultative component of a verbal compound, rather than verbal particles.

## Capturing morphological processes in Cantonese

Of particular interest to Chinese linguists is that much effort and resources in developing CANON have been put into automatic annotation of morphological processes. In addition to prefixation ( 一 'one' → 第一 'first') and suffixation ( 教 'teach' → 教師 'teacher'), there are infixation ( 討厭 'annoying' → 討鬼厭 'really annoying') and insertion of aspect marker (ASP), quantifier or verbal particle in a verbal compound ( 瞓覺 'sleep' → 瞓咗覺 'fallen asleep', 出街 'go out' → 出晒街 'all gone out', 返學 'go to school' → 返到學 'able to go to school'). The cases of infixation and insertion are particularly challenging, and automatic parsing of these processes has not been dealt with previously in a Chinese corpus, as far as we know. To illustrate, while 瞓咗覺 'fallen asleep' is treated as an insertion of the perfective ASP 咗 in the verb compound 瞓覺 in CANON, it would be marked as verb + ASP + (verb) morpheme in Wong (2006b). The latter method seems to be ad hoc and inconsistent between the verbal compound with and without ASP insertion.

Besides affixation, reduplication is a prominent morphological process in Cantonese, which may involve yes/no question formation (or A-not-A question, 驕傲 'proud' → 驕唔驕傲 'proud or not proud'), intensifying the meaning of an adjective ( 佢生得肥 'he is fat' → 佢生得肥肥 'he is very fat', or 佢做嘢不溜都穩陣 'he is always a reliable worker' → 佢做嘢不溜都穩穩陣陣 → 'he is always a very reliable worker', examples adapted from Gao (1984; pp. 60–63), adding a tentative aspect to a verb ( 睇 'to look' → 睇睇 'have a look'), a progressive aspect to a verb ( 瞓覺 'to sleep' → 瞓瞓吓覺 'in the middle of the sleep'), or a distributive property to a classifier ( 條 'classifier for long and flexible object' → 條條 'every-classifier'). These surface forms can be parsed in CANON. The morphological tags contain the underlying lexical item and the added meaning (see Table 5 for illustration). In addition, the tagger can handle situations where two morphological processes have taken place, for instance, prefixation and suffixation: 兔 ‘rabbit’ → 兔仔 ‘rabbit + diminutive’ → 阿兔仔 ‘endearing + rabbit + diminutive’, prefixation and reduplication: 妹 → 妹妹 ‘sister’ → 小妹妹 ‘little + sister’, 掉轉 ‘turn around’ → 掉轉頭 ‘turn around + suffix’ → 掉翻轉頭 ‘turn around + aspect marker insertion + suffix’.

## Annotating spoken output of PWA

One of the greatest challenges in annotating aphasic speech output corpora is the potential disagreement of parsing POS in non-fluent or grammatically ill-formed sentences of PWA. As detailed in Kong (2016), unifying linguistic information, paralinguistic aspects of oral production, and non-verbal skills co-occurring in spoken narratives can be a daunting task due to the complexity of this phenomenon, this is especially the case when an output is produced by language-impaired people. The availability of audio and video files that are corresponding to the spoken content of PWA greatly enhanced the inter- and/or intra-rater consistency of disambiguation process of annotation.

## Potential contribution to research in linguistic, psycholinguistic and neurolinguistics

The rich linguistic and prosodic data as well as videos with information on non-verbal behaviors extracted from Cantonese AphasiaBank have been proven to be extremely valuable for conducting linguistic, psycholinguistic, and neurolinguistics research in Chinese. MacWhinney and Fromm (2016) in a recent review article summarized how the AphasiaBank in English facilitated 45 research investigations in various areas of aphasic productions: (1) lexical, grammatical, and discourse output, (2) fluency and syndrome classification of aphasia, and (3) gesture employment. In addition, effects of social factors of aphasia, such as PWA’s educational background, age, gender, and occupational status, as well as intervention effects on spoken language have been explored. Here we illustrate how the Cantonese AphasiaBank may benefit researchers of different interests:

**a.** Psycholinguists in recent years have been concerned with the neural mechanisms underlying language processing and representation. Compared with the corpus that only contain natural conversational speech output of four Mandarin speakers with aphasia (Packard, 1993), our database is of a much larger scale in terms of sample size and includes a wide range of PWA with different syndromes (or types). Each participant also produced samples spanning several discourse genres arguably with different processing demands. How linguistic performance and non-verbal behaviors would vary as a function of variables such as Cantonese-speaking PWA’s age, gender, education level, or aphasia severity can be systematically analyzed. The fact that Cantonese AphasiaBank features unimpaired language dataset would allow users to compare performance between PWA and controls. For example, when contrasting the gestures of PWA and those of individuals without aphasia, videos and corresponding language samples with gesture annotations of age- and education-matched pairs of PWA and control participants can first be extracted from the database. Subsequent analyses in terms of the quantity of gestures and how their forms and functions are affected by aphasia can be conducted; this would allow a better understanding of whether and how speakers may employ gestures to enrich spoken discourse. A recent study utilizing a subset of the data bank to answer the above questions has been reported (Kong, Law, Wat, & Lai, 2015).

**b.** Neurolinguists, on the other hand, may be interested in examining how neural substrates are associated with performance in language production and comprehension. The various properties of Cantonese AphasiaBank as described above, such as lesion data of PWA, co-verbal gestures, as well as discoursal dysfluencies of repetitions, false starts, fillers, or self-corrections, constitute a rich source of information for studying theoretical processes of language production. Examples of possible questions include the impact of planning time on online language performance or effects of task demand on discourse performance (e.g., Lai, Law, & Kong, 2017).

**c.** Linguists have traditionally examined issues addressing language behaviors at specific linguistic levels, such as phonology, morphology, syntax, semantic, or discourse. Recent years have promoted an approach of language quantification that analyzes performances across linguistic levels within an individual and how they may relate to one another, namely multi-level analyses (e.g., Marini, Andreetta, Del Tin, & Carlomagno, 2011; Milman, Vega-Mendoza, & Clendenen, 2014) and multi-modal production (Linnik, Bastiaanse, & Höhle, 2016) of healthy and PWA speakers. Cantonese AphasiaBank has been utilized in a series of studies looking at issues specific to Cantonese aphasia from word to discourse levels. The major findings are summarized in Table 6. The application of multi-level analyses to data in Cantonese AphasiaBank is currently under way.

We have demonstrated how Cantonese AphasiaBank is instrumental to enhancing our knowledge of Chinese aphasia. The availability of language materials produced by healthy as well as impaired speakers will render various researchers the opportunity to perform systematic group-based comparisons that has not been possible until now. Further exploration of the data in Cantonese AphasiaBank by researchers, instructors, speech-language pathologists, related healthcare professionals, and student clinicians may facilitate the development of fundamental principles for managing Chinese aphasia now and in the future.

## Conclusions

This paper has described the background and processes involved in developing Cantonese AphasiaBank. The large scale database is distinguished from other existing spoken Chinese corpora in terms of the availability of spoken discourse and co-verbal gestures by speakers with and without aphasia containing verbal and non-verbal annotations. It promotes the use of open-source web-based access corpora for studying spoken language in a variety of discourse types among native Cantonese speakers and PWA. The rich information from the database has led to a series of linguistic, psycholinguistic and neurolingusitic studies. Further contribution of the corpora towards research and educational and/or clinical use will depend on its wider usage by researchers across disciplines.

## Acknowledgments

This study is supported by a grant funded by the National Institutes of Health to Anthony Pak-Hin Kong (PI) and Sam-Po Law (Co-I) (project number: NIH-R01-DC010398). Special thanks to the staff members in the following organizations (in alphabetical order) for their help in subject recruitment: Christian Family Service Center (Kwun Tong Community Rehabilitation Day Center), Community Rehabilitation Network of The Hong Kong Society for Rehabilitation, Internal Aphasia Clinic at the University of Hong Kong, Hong Kong Stroke Association, Lee Quo Wei Day Rehabilitation and Care Centre of The Hong Kong Society for Rehabilitation, and Self Help Group for the Brain Damage. The authors would also like to acknowledge the invaluable contribution of the people living with aphasia who participated in this study.

## References

Chafe, W. The pear stories: Cognitive, cultural and linguistic aspects of narrative production. Norwood, NJ: Ablex; 1980.

Chou, FC; Tseng, CY. The design of prosodically oriented Mandarin speech database. Proceedings of the 17th International Congress on Phonetic Sciences; 1999. 2375–2377.

Chui K, Lai HL. 2008; The NCUU Corpus of Spoken Chinese: Mandarin, Hakka, and Southern Min. Taiwan Journal of Linguistics. 6(2):119–144.

Faroqi-Shah Y. 2016; The rise of big data in neurorehabilitation. Seminars in Speech and Language. 37:3–9. [PubMed: 26882360]

Fletcher, P, Leung, C-SS, Stokes, S, Weizman, Z. Research project entitled “Milestones in the learning of spoken Cantonese by pre-school Children” funded by the Language Fund of Hong Kong. 2000. Cantonese pre-school language development: A guide.

Gāo, NH. Investigations of Cantonese dialects. Hong Kong: Hong Kong Commercial Press; 1984. Guǎng zhōu fāng y án yán jiū.

Hong Kong SAR Government. Report of the task force on population policy. 2003. Available at http://www.info.gov.hk/info/population/eng/

Kim JS. 2014; Stroke in Asia: A global disaster. International Journal of Stroke. 9(7):856–857. DOI: 10.1111/ijs.12317 [PubMed: 25231579]

Kong, APH. Analysis of neurogenic disordered discourse production: From theory to practice. New York, NY: Routledge Psychology Press; 2016.

Kong APH. 2017; Speech-Language services for Chinese-speaking people with aphasia (C-PWA): Considerations for assessment and intervention. SIG 2 Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders. 2(3):100–109. DOI: 10.1044/persp2.SIG2.100

Kong, APH, Law, S-P. Toward a multi-modal and multi-level analysis of Chinese aphasic discourse. National Institutes of Health; 2010–2014. [1-R01-DC010398]. https://projectreporter.nih.gov/project_info_description.cfm?aid=8469747

Kong APH, Law S-P, Chak G. 2015; An investigation of the use of co-verbal gestures in oral discourse among Chinese speakers with fluent versus non-fluent aphasia and healthy adults. Frontiers in Psychology. doi: 10.3389/conf.fpsyg.2015.65.00079

Kong APH, Law SP, Chak GWC. 2017; A comparison of co-verbal gesture use in oral discourse among speakers with fluent and non-fluent aphasia. Journal of Speech, Language, and Hearing Research. 60:2031–2046. DOI: 10.1044/2017_JSLHR-L-16-0093

Kong APH, Law SP, Kwan CCY, Lai C, Lam V. 2015; A coding system with independent annotations of gesture forms and functions during verbal communication: Development of a Database of Speech and GEsture (DoSaGE). Journal of Nonverbal Behavior. 39(1):93–111. DOI: 10.1007/s10919-014-0200-6 [PubMed: 25667563]

Kong APH, Law SP, Lee ASY. 2010; An investigation of use of non-verbal behaviors among individuals with aphasia in Hong Kong: Preliminary data. Procedia Social and Behavioral Sciences. 6:57–58.

Kong APH, Law SP, Wat WKC, Lai C. 2015; Co-verbal gestures among speakers with aphasia: Influence of aphasia severity, linguistic and semantic skills, and hemiplegia on gesture employment in oral discourse. Journal of Communication Disorders. 56:88–102. DOI: 10.1016/j.jcomdis.2015.06.007 [PubMed: 26186256]

Kong APH, Linnik A, Law S, Shum W. 2014; Measuring the coherence of healthy and aphasic discourse production in Chinese using Rhetorical Structure Theory (RST). Frontiers in Psychology. doi: 10.3389/conf.fpsyg.2014.64.00028

Kong APH, Linnik A, Law S, Shum W. 2017; Measuring discourse coherence in anomic aphasia using Rhetorical Structure Theory. International Journal of Speech-Language Pathology. doi: 10.1080/17549507.2017.1293158

Lausberg H, Sloetjes H. 2009; Coding gestural behavior with the NEUROGES-ELAN system. Behavior Research Methods. 41(3):841–844. [PubMed: 19587200]

Lai CCT, Law SP, Kong APH. 2017; A quantitative study of right dislocation in Cantonese spoken discourse. Language and Speech. 60(4):633–642. DOI: 10.1177/0023830916688028 [PubMed: 28139167]

Law S-P, Kong APH, Lai C. 2017; An analysis of topics and vocabulary in Chinese oral narratives by normal speakers and speakers with fluent aphasia. Clinical Linguistics and Phonetics. doi: 10.1080/02699206.2017.1334092

Law SP, Kong APH, Lai LWS, Lai C. 2015; Effects of context and word class on lexical retrieval in Chinese speakers with anomic aphasia. Aphasiology. 29(1):81–100. DOI: 10.1080/02687038.2014.951598 [PubMed: 25505810]

Lee, J. Toward a Parallel Corpus of Spoken Cantonese and Written Chinese. Proceedings of the 5th International Joint Conference on Natural Language Processing; 2011. 1462–1466.

Lee T, Kong APH, Chan VCF, Wang H. 2013; Analysis of auto-aligned and auto-segmented oral discourse by speakers with aphasia: A preliminary study on the acoustic parameter of duration. Procedia Social and Behavioral Sciences. 94:71–72.

Lee T, Kong APH, Wang H. 2014; Duration of content and function words in oral discourse by speakers with fluent aphasia: Preliminary data. Frontiers in Psychology. doi: 10.3389/conf.fpsyg.2014.64.00039

Lee T, Lo WK, Ching PC, Meng H. 2002; Spoken language resources for Cantonese speech processing. Speech Communication. 36:327–342.

Lee, T; Lam, WK; Kong, APH; Law, S-P. Analysis of intonation patterns in Cantonese aphasia speech. Oral presentation at the 18th Oriental Chapter of International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA)/Conference on Asian Spoken Language Research and Evaluation (CASLRE); Shanghai, China. 2015 Oct.

Lee T, Wong C. 1998; CANCORP: The Hong Kong Cantonese Child Language Corpus. Cahiers de Linguistique – Asie Orientale. 27(2):211–228.

Leung MT, Law SP. 2001; HKCAC: The Hong Kong Cantonese Adult Language Corpus. International Journal of Corpus Linguistics. 6(2):135–158.

Leung MT, Law SP, Fung SY. 2004; Type and token frequencies of phonological units in Hong Kong Cantonese. Behavior Research Methods, Instruments, & Computers. 36(3):500–505.

Linguistic Data Consortium. 1992. http://www.ldc.upenn.edu/

Linnik A, Bastiaanse R, Höhle B. 2016; Discourse production in aphasia: A current review of theoretical and methodological challenges. Aphasiology. 30(7):765–800. DOI: 10.1080/02687038.2015.1113489

Luke, K-K; Nancarrow, O. Hong Kong University Cantonese corpus. 1997. Available at: http://www0.hku.hk/hkcancor/

Lyle RC. 1981; A performance test for assessment of upper limb function in physical rehabilitation treatment and research. International Journal of Rehabilitation. 4(4):483–492.

MacWhinney, B. The CHILDES project: Tools for analyzing talk. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.

MacWhinney, B. The CHILDES project: Tools for analyzing talk. Hillsdale, NJ: Lawrence Erlbaum; 2003.

MacWhinney, B. Morphosyntactic analysis of the CHILDES and TalkBank corpora. In: Calzolari, N, Choukri, K, editors. Proceedings of the Eight International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA); 2012. 2375–2380. http://www.lrec-conf.org/proceedings/lrec2012/summaries/616.html

MacWhinney B, Fromm D. 2016; AphasiaBank as BigData. Seminars in Speech and Language. 37(1): 10–22. DOI: 10.1055/s-0036-1571357 [PubMed: 26882361]

MacWhinney B, Fromm D, Forbes M, Holland A. 2011; AphasiaBank: Methods for studying discourse. Aphasiology. 25:1286–1307. [PubMed: 22923879]

MacWhinney, B; Holland, A; Forbes, M; Spector, L; Fromm, D. AphasiaBank: An international archive of aphasic language. Paper presented at Meeting of Computer-based Intervention and Diagnostic Procedure- Applications for Language Impaired Persons; Vienna, Austria. 2008.

Marini A, Andreetta S, Del Tin S, Carlomagno S. 2011; A multi-level approach to the analysis of narrative language in aphasia. Aphasiology. 25:1372–1392.

Martin N, Kalinyak-Fliszar M. 2014; The case for single-case studies in treatment research—comments on Howard, Best and Nickels “Optimising the design of intervention studies: critiques and ways forward”. Aphasiology. 29(5):570–574. DOI: 10.1080/02687038.2014.987049 [PubMed: 28824217]

Milman L, Vega-Mendoza M, Clendenen D. 2014; Integrated training for aphasia: An application of part–whole learning to treat lexical retrieval, sentence production, and discourse-level communications in three cases of nonfluent aphasia. American Journal of Speech-Language Pathology. 23:105–119. [PubMed: 24686892]

Packard, JL. A linguistic investigation of aphasic Chinese speech. Dordrecht: Kluwer Academic Publishers; 1993.

Salter K, Teasell R, Foley N, Allen L. 2013The evidence-based review of stroke rehabilitation (EBRSR) reviews current practices in stroke rehabilitation. EBRSR.com.

Tardif, T. Unpublished doctoral dissertation. Yale University; 1993. Adult-to-child speech and language acquisition in Mandarin Chinese.

Wade DT, Hewer RL, David RM, Menderby PM. 1986; Aphasia after stroke: Natural history and associated deficits. Journal of Neurology, Neurosurgery & Psychiatry. 49:11–16.

Wang J. 2001; Recent progress in corpus linguistics in China. International Journal of Corpus Linguistics. 6(2):281–304.

Willmes K. 2007; Statistical methods for a single-case study approach to aphasia therapy research. Aphasiology. 4(4):415–436. DOI: 10.1080/02687039008249092

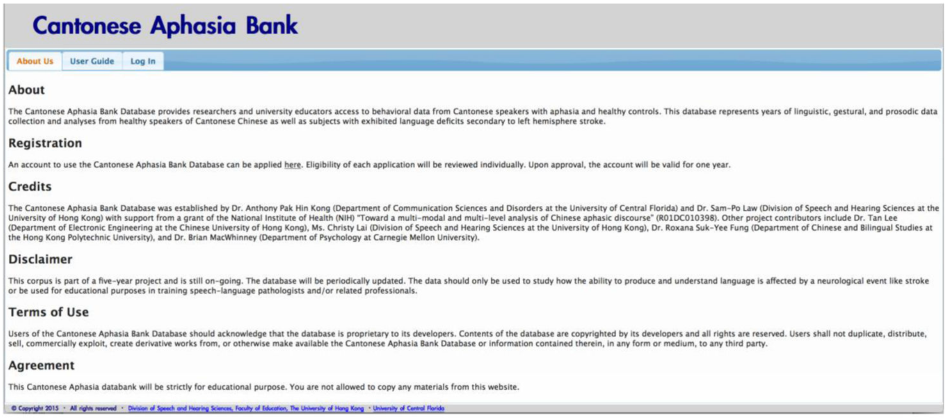
Wong PW. 2006a; The specification of POS tagging of the Hong Kong University Cantonese Corpus. International Journal of Technology and Human Interaction. 2(1):21–38.

Wong, PW-Y. Unpublished doctoral dissertation. Ohio State University; 2006b. Syllable fusion in Hong Kong Cantonese connected speech.

Xu, L-J, Lee, T. Parametric variation in three Chinese dialects, Cantonese, Shanghainese and Mandarin. Hong Kong: Research Grant Council; 1998.

Yang, X-J. Survey and prospect of China’s corpus-based research. In: Wilson, A, Archer, D, Rayson, P, editors. Corpus linguistics around the world. New York: Rodopi; 2006. 219–233.

Zhou, K; Li, A; Yin, Z; Zong, C. CASIA-CASSIL: A Chinese telephone conversation corpus in real scenarios with multi-leveled annotation. The 7th International Conference on Language Resources and Evaluation; 2010. 2407–2413.



**About**

The Cantonese Aphasia Bank Database provides researchers and university educators access to behavioral data from Cantonese speakers with aphasia and healthy controls. This database represents years of linguistic, gestural, and prosodic data collection and analyses from healthy speakers of Cantonese Chinese as well as subjects with exhibited language deficits secondary to left hemisphere stroke.

**Registration**

An account to use the Cantonese Aphasia Bank Database can be applied here. Eligibility of each application will be reviewed individually. Upon approval, the account will be valid for one year.

**Credits**

The Cantonese Aphasia Bank Database was established by Dr. Anthony Pak Hin Kong (Department of Communication Sciences and Disorders at the University of Central Florida) and Dr. Sam-Po Law (Division of Speech and Hearing Sciences at the University of Hong Kong) with support from a grant of the National Institute of Health (NIH) "Toward a multi-modal and multi-level analysis of Chinese aphasic discourse" (R01DC010398). Other project contributors include Dr. Tan Lee (Department of Electronic Engineering at the Chinese University of Hong Kong), Ms. Christy Lai (Division of Speech and Hearing Sciences at the University of Hong Kong), Dr. Roxana Suk-Yee Fung (Department of Chinese and Bilingual Studies at the Hong Kong Polytechnic University), and Dr. Brian MacWhinney (Department of Psychology at Carnegie Mellon University).

**Disclaimer**

This corpus is part of a five-year project and is still on-going. The database will be periodically updated. The data should only be used to study how the ability to produce and understand language is affected by a neurological event like stroke or be used for educational purposes in training speech-language pathologists and/or related professionals.

**Terms of Use**

Users of the Cantonese Aphasia Bank Database should acknowledge that the database is proprietary to its developers. Contents of the database are copyrighted by its developers and all rights are reserved. Users shall not duplicate, distribute, sell, commercially exploit, create derivative works from, or otherwise make available the Cantonese Aphasia Bank Database or information contained therein, in any form or medium, to any third party.

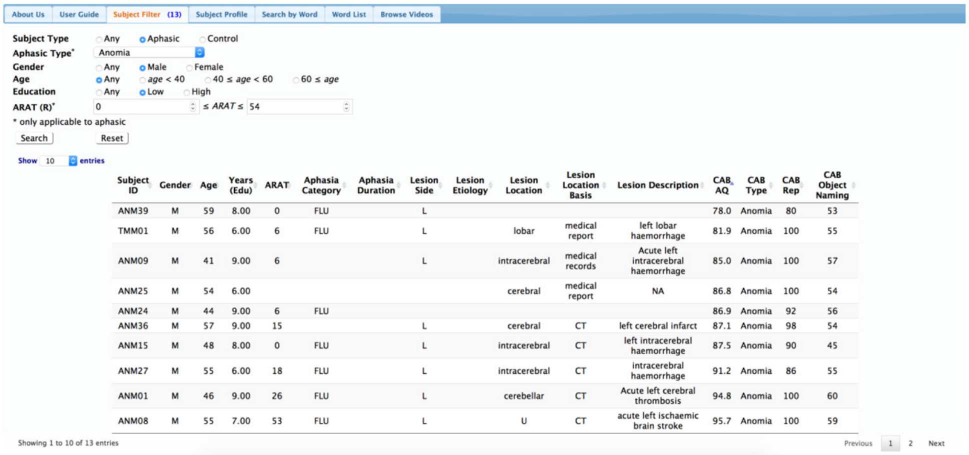
**Agreement**

This Cantonese Aphasia databank will be strictly for educational purpose. You are not allowed to copy any materials from this website.

© Copyright 2015 · All rights reserved · Division of Speech and Hearing Sciences, Faculty of Education, The University of Hong Kong · University of Central Florida

**Figure 1.**

The Cantonese AphasiaBank 'About us' and 'Log in' screenshot



Subject Type ○ Any ● Aphasic ○ Control

Aphasic Type* Anomia

Gender ○ Any ● Male ○ Female

Age ● Any ○ *age* < 40 ○ 40 ≤ *age* < 60 ○ 60 ≤ *age*

Education ○ Any ● Low ○ High

ARAT (R)* 0 ≤ *ARAT* ≤ 54

* only applicable to aphasic

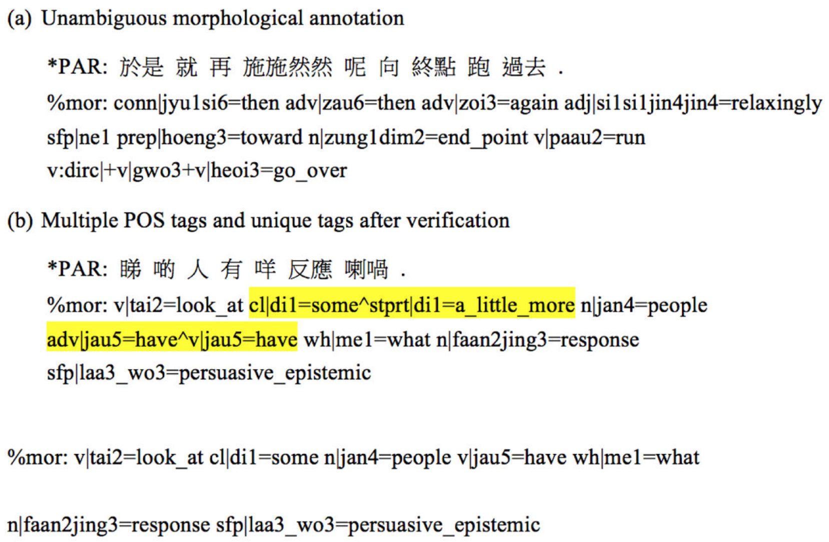
Show 10 entries

| Subject ID | Gender | Age | Years (Edu) | ARAT | Aphasia Category | Aphasia Duration | Lesion Side | Lesion Etiology | Lesion Location | Lesion Location Basis | Lesion Description | CAB AQ | CAB Type | CAB Rep | CAB Object Naming |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ANM39 | M | 59 | 8.00 | 0 | FLU |  | L |  |  |  |  | 78.0 | Anomia | 80 | 53 |
| TMM01 | M | 56 | 6.00 | 6 | FLU |  | L |  | lobar | medical report | left lobar haemorrhage | 81.9 | Anomia | 100 | 55 |
| ANM09 | M | 41 | 9.00 | 6 |  |  | L |  | intracerebral | medical records | Acute left intracerebral haemorrhage | 85.0 | Anomia | 100 | 57 |
| ANM25 | M | 54 | 6.00 |  |  |  |  |  | cerebral | medical report | NA | 86.8 | Anomia | 100 | 54 |
| ANM24 | M | 44 | 9.00 | 6 | FLU |  |  |  |  |  |  | 86.9 | Anomia | 92 | 56 |
| ANM36 | M | 57 | 9.00 | 15 |  |  | L |  | cerebral | CT | left cerebral infarct | 87.1 | Anomia | 98 | 54 |
| ANM15 | M | 48 | 8.00 | 0 | FLU |  | L |  | intracerebral | CT | left intracerebral haemorrhage | 87.5 | Anomia | 90 | 45 |
| ANM27 | M | 55 | 6.00 | 18 | FLU |  | L |  | intracerebral | CT | intracerebral haemorrhage | 91.2 | Anomia | 86 | 55 |
| ANM01 | M | 46 | 9.00 | 26 | FLU |  | L |  | cerebellar | CT | Acute left cerebral thrombosis | 94.8 | Anomia | 100 | 60 |
| ANM08 | M | 55 | 7.00 | 53 | FLU |  | L |  | U | CT | acute left ischaemic brain stroke | 95.7 | Anomia | 100 | 59 |

Showing 1 to 10 of 13 entries

**Figure 2.**

A screenshot of subject filter features and results in Cantonese AphasiaBank



(a) Unambiguous morphological annotation

*PAR: 於是 就 再 施施然然 呢 向 終點 跑 過去 .
%mor: conn|jyu1si6=then adv|zau6=then adv|zoi3=again adj|si1si1jin4jin4=relaxingly sfp|ne1 prep|hoeng3=toward n|zung1dim2=end_point v|paau2=run v:dirc|+v|gwo3+v|heoi3=go_over

(b) Multiple POS tags and unique tags after verification

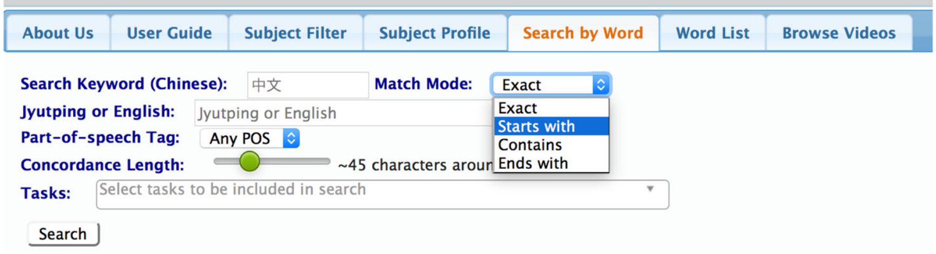
*PAR: 睇 啲 人 有 咩 反應 喇喎 .
%mor: v|tai2=look_at cl|di1=some^stprt|di1=a_little_more n|jan4=people adv|jau5=have^v|jau5=have wh|me1=what n|faan2jing3=response sfp|laa3_wo3=persuasive_epistemic

%mor: v|tai2=look_at cl|di1=some n|jan4=people v|jau5=have wh|me1=what

n|faan2jing3=response sfp|laa3_wo3=persuasive_epistemic

**Figure 3.**

Format and levels of annotation



Search Keyword (Chinese): 中文 Match Mode: Exact

Jyutping or English: Jyutping or English

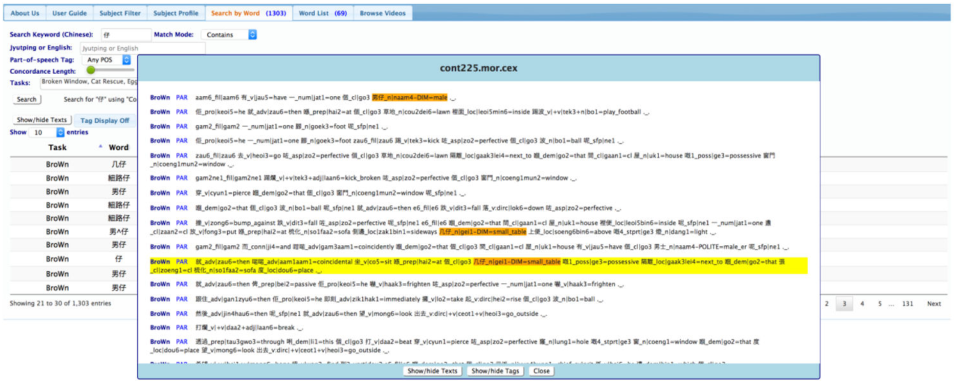
Part-of-speech Tag: Any POS

Concordance Length: ~45 characters aroun

Tasks: Select tasks to be included in search

**Figure 4.**

A screenshot of “Search by Word” filter parameters in Cantonese AphasiaBank



Search Keyword (Chinese): 仔 Match Mode: Contains

Jyutping or English: Jyutping or English

Part-of-speech Tag: Any POS

Concordance Length:

Tasks: Broken Window, Cat Rescue, Egg

Search for "仔" using "Co

Show 10 entries

| Task | Word |
|---|---|
| BroWn | 几仔 |
| BroWn | 細路仔 |
| BroWn | 男仔 |
| BroWn | 細路仔 |
| BroWn | 細路仔 |
| BroWn | 男^仔 |
| BroWn | 男仔 |
| BroWn | 仔 |
| BroWn | 男仔 |
| BroWn | 男仔 |

Showing 21 to 30 of 1,303 entries

2 3 4 5 ... 131 Next

cont225.mor.cex

BroWn PAR aam6_fil|aam6 有_v|jau5=have 一_num|jat1=one 個_cl|go3 男仔_n|naam4-DIM=male ._.

BroWn PAR 佢_pro|keoi5=he 就_adv|zau6=then 喺_prep|hai2=at 個_cl|go3 草地_n|cou2dei6=lawn 裡面_loc|leoi5min6=inside 踢波_v|+v|tek3+n|bo1=play_football ._.

BroWn PAR gam2_fil|gam2 一_num|jat1=one 腳_n|goek3=foot 呢_sfp|ne1 ._.

BroWn PAR 佢_pro|keoi5=he 一_num|jat1=one 腳_n|goek3=foot zau6_fil|zau6 踢_v|tek3=kick 咗_asp|zo2=perfective 個_cl|go3 波_n|bo1=ball 呢_sfp|ne1 ._.

BroWn PAR zau6_fil|zau6 去_v|heoi3=go 咗_asp|zo2=perfective 個_cl|go3 草地_n|cou2dei6=lawn 隔籬_loc|gaak3lei4=next_to 嗰_dem|go2=that 間_cl|gaan1=cl 屋_n|uk1=house 嘅1_poss|ge3=possessive 窗門_n|coeng1mun2=window ._.

BroWn PAR gam2ne1_fil|gam2ne1 踢爛_v|+v|tek3+adj|laan6=kick_broken 咗_asp|zo2=perfective 個_cl|go3 窗門_n|coeng1mun2=window ._.

BroWn PAR 穿_v|cyun1=pierce 嗰_dem|go2=that 個_cl|go3 窗門_n|coeng1mun2=window 呢_sfp|ne1 ._.

BroWn PAR 嗰_dem|go2=that 個_cl|go3 波_n|bo1=ball 呢_sfp|ne1 就_adv|zau6=then e6_fil|e6 跌_v|dit3=fall 落_v:dirc|lok6=down 咗_asp|zo2=perfective ._.

BroWn PAR 撞_v|zong6=bump_against 跌_v|dit3=fall 咗_asp|zo2=perfective 呢_sfp|ne1 e6_fil|e6 嗰_dem|go2=that 間_cl|gaan1=cl 屋_n|uk1=house 裡便_loc|leoi5bin6=inside 呢_sfp|ne1 一_num|jat1=one 盞_cl|zaan2=cl 放_v|fong3=put 喺_prep|hai2=at 梳化_n|so1faa2=sofa 側邊_loc|zak1bin1=sideways 几仔_n|gei1-DIM=small_table 上便_loc|soeng6bin6=above 嘅4_stprt|ge3 燈_n|dang1=light ._.

BroWn PAR gam2_fil|gam2 而_conn|ji4=and 咁啱_adv|gam3aam1=coincidently 嗰_dem|go2=that 個_cl|go3 間_cl|gaan1=cl 屋_n|uk1=house 有_v|jau5=have 個_cl|go3 男士_n|naam4-POLITE=male_er 呢_sfp|ne1 ._.

BroWn PAR 就_adv|zau6=then 啱啱_adv|aam1aam1=coincidental 坐_v|co5=sit 喺_prep|hai2=at 個_cl|go3 几仔_n|gei1-DIM=small_table 嘅1_poss|ge3=possessive 隔籬_loc|gaak3lei4=next_to 嗰_dem|go2=that 張_cl|zoeng1=cl 梳化_n|so1faa2=sofa 度_loc|dou6=place ._.

BroWn PAR 就_adv|zau6=then 俾_prep|bei2=passive 佢_pro|keoi5=he 嚇_v|haak3=frighten 咗_asp|zo2=perfective 一_num|jat1=one 嚇_v|haak3=frighten ._.

BroWn PAR 跟住_adv|gan1zyu6=then 佢_pro|keoi5=he 即刻_adv|zik1hak1=immediately 攞_v|lo2=take 起_v:dirc|hei2=rise 個_cl|go3 波_n|bo1=ball ._.

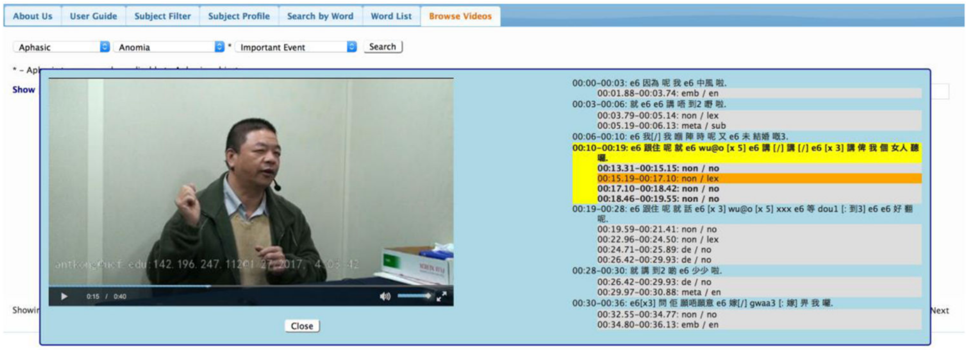
BroWn PAR 然後_adv|jin4hau6=then 呢_sfp|ne1 就_adv|zau6=then 望_v|mong6=look 出去_v:dirc|+v|ceot1+v|heoi3=go_outside ._.

BroWn PAR 打爛_v|+v|daa2+adj|laan6=break ._.

BroWn PAR 透過_prep|tau3gwo3=through 呢_dem|li1=this 個_cl|go3 打_v|daa2=beat 穿_v|cyun1=pierce 咗_asp|zo2=perfective 窿_n|lung1=hole 嘅4_stprt|ge3 窗_n|coeng1=window 嗰_dem|go2=that 度_loc|dou6=place 望_v|mong6=look 出去_v:dirc|+v|ceot1+v|heoi3=go_outside ._.

**Figure 5.**

A screenshot of “Search by Word” results in Cantonese AphasiaBank with corresponding keywords highlighted. The code ‘BroWn’ in the text box indicates that these sentence lines were found in the sequential picture description task of the “Broken Window” drawings.



```
00:00–00:03: e6 因為 呢 我 e6 中風 啦.
        00:01.88–00:03.74: emb / en
00:03–00:06: 就 e6 e6 講 唔 到2 嘢 啦.
        00:03.79–00:05.14: non / lex
        00:05.19–00:06.13: meta / sub
00:06–00:10: e6 我[/] 我 嗰 陣 時 呢 又 e6 未 結婚 嘅3.
00:10–00:19: e6 跟住 呢 就 e6 wu@o [x 5] e6 講 [/] 講 [/] e6 [x 3] 講 俾 我 個 女人 聽
        囉.
        00:13.31–00:15.15: non / no
        00:15.19–00:17.10: non / lex
        00:17.10–00:18.42: non / no
        00:18.46–00:19.55: non / no
00:19–00:28: e6 跟住 呢 就 話 e6 [x 3] wu@o [x 5] xxx e6 等 dou1 [: 到3] e6 e6 好 翻
        呢.
        00:19.59–00:21.41: non / no
        00:22.96–00:24.50: non / lex
        00:24.71–00:25.89: de / no
        00:26.42–00:29.93: de / no
00:28–00:30: 就 講 到2 啲 e6 少少 啦.
        00:26.42–00:29.93: de / no
        00:29.97–00:30.88: meta / en
00:30–00:36: e6[x3] 問 佢 願唔願意 e6 嫁[/] gwaa3 [: 嫁] 畀 我 囉.
        00:32.55–00:34.77: non / no
        00:34.80–00:36.13: emb / en
```

**Figure 6.**

A screenshot of viewing recordings of subjects in Cantonese AphasiaBank with synchronized display of video, language transcriptions, and gesture annotations

## Table 1

Distribution of Unimpaired Participants and Persons with Aphasia (PWA) by Age and Education Subgroups

| | **Male** | | **Female** | | |
|---|---|---|---|---|---|
| **Unimpaired Participants** | | | | | |
| **Age (in year)** | Low Education | High Education | Low Education | High Education | **Sub-total** |
| **18 – 39** | 12 | 12 | 10 | 13 | 47 |
| **40 – 59** | 15 | 9 | 12 | 12 | 48 |
| **60 or above** | 11 | 15 | 22 | 6 | 54 |
| **Sub-total** | 38 | 36 | 44 | 31 | **(149)** |
| Persons with Aphasia | | | | | |
| **18 – 39** | 3 (1 Wernicke's + 1 Broca's + 1 Global) | 2 (1 Transcortical motor + 1 Isolation) | 2 (2 Anomic) | 1 (1 Transcortical motor) | 8 |
| **40 – 59** | 47 (26 Anomic + 4 Transcortical sensory + 11 Broca's + 5 Transcortical motor + 1 Global) | 5 (4 Anomic + 1 Transcortical sensory) | 15 (11 Anomic + 1 Broca's + 3 Transcortical motor) | 1 (1 Transcortical motor) | 68 |
| **60 or above** | 10 (5 Anomic + 1 Transcortical sensory + 1 Broca's + 3 Transcortical motor) | 3 (3 Anomic) | 10 (5 Anomic + 1 Wernicke's + 2 Broca's + 1 Transcortical motor + 1 Isolation) | 6 (4 Anomic + 1 Broca's + 1 Global) | 29 |
| **Sub-total** | 60 | 10 | 27 | 8 | **(105)** |

Note. Higher or lower than secondary school for the two younger groups, and higher or low than primary school for the oldest group, i.e., at least 14 years as higher education for the two younger groups and at least 7 years as higher education for the elderly group. The high illiteracy rate in native Cantonese elderly speakers in Hong Kong (and in Mainland China as well as many Chinese oversea communities) limited our application of the same high-low education criterion for all participant groups. According to Hong Kong SAR Government (2003), over half of the Hong Kong population during 1970s had only attained primary school education; the percentage of was even higher prior to that. Furthermore, over the decades, many jobs now requiring tertiary education qualification previously only required secondary school education. We believe that education and working experience could associate with language abilities of our elderly group participants. This had led us to draw this equivalence (of 7 years of education for the oldest group is comparable to 14 years of education for the two younger groups, in terms of language abilities).

## Table 2

Demographic Information on Persons with Aphasia (PWA)

| | **Fluent** | | | **Non-fluent** | | | |
|---|---|---|---|---|---|---|---|
| **Aphasia syndrome** | **Anomic** | **Wernicke's** | **Transcortical sensory** | **Broca's** | **Transcortical motor** | **Isolation** | **Global** |
| **Gender** | 39M + 21F | 1M + 1F | 6M + 0F | 13M + 4F | 9M + 6F | 1M + 1F | 2M + 1F |
| **Age (in years)** | 55.00 (9.97)[a] | 52.92 (5.70) | 49.33 (15.91) | 55.24 (8.53) | 57.31 (14.21) | 50.92 (16.38) | 49.92 (13.75) |
| | 55.00 (9.97) | | | 55.41 (11.58) | | | |
| | 55.15 (10.51) | | | | | | |
| **Years of education** | 9.07 (3.55) | 5.50 (7.78) | 11.00 (3.29) | 8.71 (3.04) | 10.13 (5.54) | 9.50 (4.95) | 10.00 (2.65) |
| | 9.13 (3.67) | | | 9.43 (4.19) | | | |
| | 9.24 (3.85) | | | | | | |
| **Aphasia Quotient (AQ)**[b] | 89.84 (7.37) | 65.35 (11.10) | 73.83 (11.67) | 47.81 (13.33) | 69.89 (9.65) | 11.45 (0.64) | 44.47 (5.57) |
| | 87.70 (9.80) | | | 54.52 (18.66) | | | |
| | 76.01 (20.88) | | | | | | |
| **Post onset time (in years)** | 5.18 (4.51) | 1.58 (1.30) | 2.43 (2.08) | 4.25 (5.50) | 4.09 (3.58) | 3.63 (2.89) | 2.42 (1.52) |
| | 4.83 (4.38) | | | 4.01 (4.326) | | | |
| | 4.54 (4.37) | | | | | | |

Note.

[a] values are given in "mean and (standard deviations)".

[b] Based on the results of the Cantonese version of the Western Aphasia Battery (Yiu, 1992) and the maximum is 100, a higher score of AQ indicates a lower severity of language impairment.

M = male; F = female.

## Table 3

List of Gesture Annotations in Cantonese AphasiaBank

| Gesture annotation | Abbreviation |
|---|---|
| **Forms** | |
| Iconic | i |
| Metaphoric | meta |
| Deictic | dei |
| Emblem | em |
| Beats | be |
| Non-identifiable | non |
| **Functions** | |
| Providing additional information | pro |
| Enhancing language content | en |
| Alternate means of communication | alt |
| Guiding speech flow | gui |
| Reinforcing prosody of speech | rein |
| Assisting lexical retrieval | lex |
| Assisting sentence reconstruction | sent |
| No specific function | no |

## Table 4

List of Part-of-Speech Tags in Cantonese AphasiaBank

| Part-of-speech | Abbreviation | Example with explanation |
|---|---|---|
| **Functional morphemes** | | |
| Affix | / | 哋 ‘plural marker’ |
| Aspectual marker | asp | 過 ‘experiential marker’ |
| Classifier | cl | 群 ‘group’ |
| Communicator | co | 唔該 ‘thank you’ |
| Conjunction | conj | 同 ‘and’ |
| Connective | conn | 因為 ‘because’ |
| Demonstrative | dem | 嗰 ‘that’ |
| Determiner | det | 今 ‘this’ |
| Expression | exp | 人生路不熟 ‘feeling out of place’ |
| Filler | fil | gam2aa6 |
| Interjection | int | 哎呀 ‘ouch’ |
| Locative | loc | 上面 ‘above’ |
| Numeral | num | 零 ‘zero’ |
| Onomatopoeia | on | 𠯠 ‘a bark’ |
| Preposition | prep | 透過 ‘through’ |
| Pro-word | pro | 大家 ‘each other’ |
| Quantifier | quan | 哂 ‘all’ |
| Sentence final particle | sfp | 喎 ‘reported speech’ |
| Structural particle | stprt | 地 ‘-ly’ |
| Verbal particle | vprt | 添 ‘also’ |
| wh-word | wh | 點樣 ‘how’ |
| **Adjective** | | |
| Adjective | adj | 得意 ‘funny’ |
| Complex adjective | adj-com | 大聲 ‘loud’ |
| **Adverb** | | |
| Adverb | adv | 亦都 ‘also’ |
| Adverb of negation | adv-neg | 冇 ‘have not’ |
| **Noun** | | |
| Noun (monosyllabic) | n-mono | 橋 ‘bridge’ |
| Noun (multisyllabic) | n-multi | 天橋 ‘overpath’ |
| Proper noun | n-prop | 黄大仙 |
| **Verb** | | |
| Auxiliary verb | aux | 應該 ‘should’ |
| Directional verb | vdirc | 出嚟 ‘come out’ |

| Part-of-speech | Abbreviation | Example with explanation |
|---|---|---|
| Verb (monosyllabic) | v-mono | 掛 ‘hang’ |
| Verb (multisyllabic) | v-multi | 志在 ‘care’ |
| Verb (adj+v) | v-adj+v | 歡呼 ‘cheer’ |
| Verb (n+v) | v-n+v | 心諗 ‘think’ |
| Verb (v+adj) | v-v+adj | 整爛 ‘break’ |
| Verb (v+n) | v-v+n | 瞓覺 ‘sleep’ |
| Verb (v+v) | v-v+v | 到達 ‘arrive’ |
| Verb (v+v+n) | v-v+v+n | 瞓醒覺 ‘awake’ |

Note: adj = adjective; n = noun; v = verb.

## Table 5

Tagging of Morphological Rules

| **Surface form** | **Morphological annotation** |
|---|---|
| *Affixation* | |
| Prefixation: 第一 ‘number one’ | ORD#num\|jat1=one |
| Suffixation: 白兔仔 ‘little rabbit’ | n\|+n\|baak6to3-DIM=rabbit |
| Infixation: 瞓緊覺 ‘sleeping’ | v\|+v\|fan3+n\|gaau3&asp:gan2=sleep |
| *Reduplication* | |
| A-not-A question: 驕唔驕傲 ‘Are you proud?’ | adj\|giu1ngou6&NEG=proud |
| Intensifying: 開開心心 ‘very happy’ | adj\|hoi1sam1&INT=happy |
| Tentative: 試試 ‘give it a try’ | v\|si3&TEN=try |
| Distributive: 日日 ‘everyday’ | n\|jat6&EVERY=day |
| *Two morphological processes* | |
| 小妹妹 ‘little sister’ | DMP#n\|mui6&redup=sister |

## Table 6

Current Major Findings of Investigations Using Cantonese AphasiaBank

| **Aspects of verbal and non-verbal communication** | **Materials used** | **Major findings** |
|---|---|---|
| Lexical processing | Connected speech (3 tasks), noun and verb naming of 19 anomic PWA and controls | With balanced age-of-acquisition, familiarity and imageability, word retrieval accuracy was higher in picture naming than connected speech, but PWA did not retrieve nouns better than verbs (Law, Kong, Lai, & Lai, 2015). |
| | Connected speech (1 task) of 65 fluent PWA and controls | Analyses of vocabulary use in terms of part-of-speech, word frequency, lexical semantics, and diversity were conducted. It was found that the vocabulary size was larger for controls than that of PWA, but the distribution across parts-of-speech, frequency of occurrence, and the ratio of concrete to abstract items in major open word classes was similar for both speaker groups. In addition, the two groups proportionately used more different verbs than nouns. (Law, Kong, Lai, & Lai, 2017). |
| Micro-linguistic performance | Connected speech (3 tasks) of 144 unimpaired speakers | Genre types (i.e., structure and style) and planning load influenced the rate of occurrence of right dislocation, a focus marking device that carried an affective function motivated by limited planning time in conversation (Lai, Law, & Kong, 2017). |
| | Connected speech (3 tasks) of 131 unimpaired speakers and 48 PWA | PWA were impaired in various aspects of micro-linguistic skills, including reduced type-token ratio, lower percentage of simple or complete sentences, higher percentage of regulators and dysfluency (Kong, Law, Kwan, Lai, & Lam, 2015; Kong, Law, Wat, & Lai, 2015). |
| Co-verbal gestures | | PWA employed significantly more gestures than controls. Gesture use was influenced by aphasia severity, semantic integrity, and linguistic competence (Kong, Law, Wat, & Lai, 2015). |
| | Connected speech (3 tasks) of 23 fluent PWA, 21 non-fluent PWA, and 23 controls | Different patterns of using non-verbal behaviors were observed between the two PWA groups (Kong, Law, & Chak, 2015, 2017). |
| Non-verbal behaviors | Connected speech (2 tasks) of 2 fluent PWA, 2 non-fluent PWA, and controls | Frequency of employing referent-related gestures, functional gestures, facial expressions, and adaptors in PWA was over double than controls. Fluent and non-fluent PWA differed qualitatively in using these non-verbal behaviors (Kong, Law, & Lee, 2010). |
| Macro-linguistic performance | Connected speech (2 tasks) of 15 anomic PWA and controls | PWA were reduced in information content and elaboration with simplified/impaired discourse structures. Their macro-linguistic deficits were evidenced by impaired cohesion and coherence (Kong, Linnik, Law, & Shum, 2014, 2017). |
| | Connected speech (1 task) of 65 fluent PWA and controls | When engaged in a monologue of telling an important event of the life, there was a significant overlap in the topics between the two speaker groups (Law, Kong, Lai, & Lai, 2017). |
| Speech prosody | All narrative tasks from 17 fluent PWA and controls | PWA were impaired in prosody that contained significantly more sentences with rising intonation (more noticeable for shorter sentences) (Lee, Lam, Kong, & Law, 2015). PWA's syllable duration was significantly longer with more inappropriate speech pauses (Lee, Kong, Chan, & Wang, 2013). The durations of content versus function words in oral discourse were also different (Lee, Kong, & Wang, 2014). |

Note: PWA = Persons with aphasia.