The 3rd International Symposium on Linguistic Patterns in Spontaneous Speech (LPSS2019)
Institute of Linguistics, Academia Sinica, Taipei
November 21-22, 2019

ILAS Chinese Spoken Language Resources

Shu-Chuan Tseng

Institute of Linguistics, Academia Sinica
tsengsc@gate.sinica.edu.tw

Abstract

Spoken language resources annotated with information about specific phenomena facilitate objective empirical linguistic observations, adequate data-driven model construction, and efficient goal-oriented system development. Impactful research outcomes and application opportunities benefit explicitly and implicitly from well-annotated human speech resources. As far as spoken language resources are concerned, collecting large-scale audio files is no longer a difficult task with current technology and media. However, it is still challenging to conduct high-quality, theory-based speech annotation projects. This paper gives an overview of four spoken language resources that have been constructed at the Institute of Linguistics, Academia Sinica (ILAS), for various research projects involving phonetic-acoustic correlates of production units in adult conversation, sociophonetic indices, and child speech pathology as well as acquisition.

1. Introduction

1.1. Human speech as a complex system

The development of experimental studies of languages and the importance of human speech in society were recognized decades ago by Charles Bally [3]. In a similar way but directly pointing to phonetic studies, John Rupert Firth additionally stated that “the application of instrumental as well as of impressionist phonetics correlates with statements of phonological and grammatical meaning” [13] (p. 25). Apparently, the social function of speech is based on the recognition of human use in interactive contexts. Speech production and perception should satisfy the need for effective verbal communication. While studies of human speech as a whole take into account social interaction, conversational speech is a complex target for research. It is produced at a fast tempo with shared knowledge of the world on the one hand and individually distinctive personal experiences of the participants in a conversation on the other hand. Spontaneous speech is often reduced but clearly understandable given the shared underlying knowledge systems at multiple linguistic levels. In the twenty-first century, instrumental studies of speech are becoming deeper and broader than ever before, with a drastically growing quantity of knowledge and increasing computation power of hardware and modeling techniques. However, decoding the meaning and function of human speech is unlikely if theoretical frameworks such as phonology, grammar, semantics, and particularly pragmatics are lacking. The task of understanding speech is more than merely reporting numerical results. However, in accordance with the statements of Bally and Firth, the study of speech should be a combination of instrumental analysis, linguistic theory, and sociopragmatic accounts. Directly related to this issue, the importance of spoken language resources with respect to technological system development has attracted intense attention, particularly since the recent commencement of the data science epoch. Concerning the role of spoken resources as data materials for the fields of linguistics and speech technology, a number of impactful spoken corpora have been constructed and have already elicited many fruitful research and applications, e.g., the London Lund Corpus of English Conversation, the HCRC Map Task Corpus, the SWITCHBOARD Corpus, the Buckeye Corpus of Conversational Speech, the Corpus of Spontaneous Japanese, the Kiel Corpus of Spontaneous Speech for German, the Chinese Annotated Spontaneous Speech, the Mandarin Chinese Broadcast News Corpus and the Spoken Dutch Corpus [32] [6] [1] [14] [27] [24] [18] [21] [42] [25]. This paper will focus on four Chinese spoken language resources constructed at Academia Sinica.

1.2. Phonetic labeling in speech

Conventionally, annotation schema is designed specifically to mark up the speech phenomena to be investigated according to researcher interest. Among a wide variety of annotation works, phone boundary annotation for conversational speech that is filled with the reduction of different degrees and disfluencies is particularly challenging. Once the resources are made available [15] [27] [28] [25] [24] [30], these corpora are widely useful for linguistic analysis, as phonetic representations of reduced spoken words can then be well studied with signal-aligned phone information. They are also useful for developing sophisticated speech technology systems, e.g., pronunciation dictionary enhancement [29] [22] [33] [31]. Signal-aligned phone boundaries can be obtained by manual labeling or by automatic generation. Manual phonetic labeling requires professional training in phonetic transcription and proper knowledge of phonetics [15] [24]. Automatic generation of phone boundaries, on the other hand, requires a well-trained phone recognizer for continuous speech [29] [43] [4] [12]. Both methods of phone boundary generation have advantages and disadvantages in terms of cost and precision. Originally proposed for evaluating discourse segment annotation in text corpora, Cohen’s kappa and the follow-up issues that are related to the adequacy of statistical measures and the interrelationship between the essence of coding schema, the number of coders, the professional training of the coder (expert vs. naïve coder), etc., have led to intensive discussions [9] [2]. Given that the quality of annotation matters and it is not practical to obtain intercoder agreement for large-scale phone labeling projects, it is a common practice to evaluate only a small subset [29]. Another alternative is to manually verify the automatic results at certain selected stages of data processing [25] [27].

2. Construction of Chinese spoken corpora

2.1. Words as our elementary processing unit

Mandarin Chinese is an analytic language that rarely has inflection. Suffixes and markers are often used instead, e.g., for plural 們 *men*, past tense 過 *guo*, and temporal aspect 了 *le* [7]. Word order is thus essential in interpreting the meaning of sentences. In addition, with the large number of compound words in Chinese, the co-occurrence of words and syllables is very relevant when decoding the semantic meaning. A Chinese syllable, very often a morpheme at the same time, is written in terms of a character that conveys not only the meaning of the syllable but also the pronunciation, including the distinction in lexical tones. In ancient Chinese, the majority of words are monosyllables [41], in contrast with modern Chinese, which consists of a larger number of disyllabic word types [38]. In particular, grammatical words in Chinese are generally monosyllabic. Chinese is also a language with lexical tones. It is mainly spoken in China and Taiwan. Though it has distinctive prosodic patterns (e.g., intonation and stressed vs. unstressed syllables), writing systems (simplified vs. traditional characters), and lexical choices in specific domains, the sound inventory and tone system [10] [16] are overall identical in the two variants, with subtle discrepancies due to regional or idiolect differences. Please note that the terms Chinese and Taiwan Mandarin used in this paper are meant to be the variant of Mandarin Chinese spoken in Taiwan.

In our corpus construction projects, we have a fundamental assumption regarding the role of "words". In speech production models, the meaning and the (phonological) form of words are stored in the mental lexicon, functioning as building blocks for processing and producing higher levels of linguistic units, e.g., phrase, clause, and sentence [20]. For automatic speech recognition systems, the pronunciation dictionary consists of shortlisted candidates of phone sequences representing the canonical forms and, in certain cases, also pronunciation variants of spoken words [17]. Because a word is a semantic unit, we assume a reasonable degree of invariance between the meaning and the form of words to be preserved. This is similar to the notion of the exemplar theory, in which "each category is represented in memory by a large cloud of remembered tokens of that category. These memories are organized in a cognitive map, so that memories of highly similar instances are close to each other and memories of dissimilar instances are far apart. The remembered tokens display the range of variation that is exhibited in the physical manifestations of the category" [26] (p. 140).

Thus, in our hybrid approach, human verification is conducted at the word level only. In particular, the labeling of lower levels of units, such as phones and syllables, requires intensive attention. On the other hand, phrases or clauses are rather long. It would not be an adequate working unit if we aim to minimize the misalignment of phone boundaries by automatic aligners, often due to the high degree of variability caused by prosodic properties and phonetic reduction, e.g., syllable omission. As a whole, word boundaries can be relatively easily perceived and judged. This accelerates the verification process and guarantees a certain degree of consistency across labelers. In particular for Chinese, reduced or omitted syllable boundaries in highly contracted disyllabic words are usually blurred [37]. It is difficult not only for systems but also for human labelers to locate the position of omitted boundaries in a consistent way. In our working schema, the labelers only need to pay attention to the perceptual adequacy of word boundaries.

2.2. Constructing multilayer linguistic annotation

Our spoken resource projects are in principle concerned with continuous speech, including adult conversational speech, adult interview speech, child repetitive speech, and child narrative speech. The only exception is the child phonological development speech corpus, in which spoken words are recorded in isolation and manual verification is conducted at the boundaries of syllables. For details about orthographic transcription conventions, e.g., Chinese characters, discourse markers, particles, fillers, paralinguistic sounds, and code-switching, please refer to [38]. Long speech stretches are segmented into interpause units (IPUs) according to disjuncture cues such as pauses and paralinguistic sounds (e.g., breathing, inhalation, and laughter). The ILAS phone aligner is applied to automatically obtain boundary information on phonemes and syllables based on the text information from transcripts. The performance of the ILAS phone aligner was empirically evaluated by three pairs of professional labelers working on 5,276 IPUs of speech data. The deviation of the phone boundary within each pair of labelers was reported to be less than 20 milliseconds in more than 90% of the test data [23].

Signal-aligned word boundaries are derived by making use of syllable boundary information (signal-based) and word segmentation results (text-based). Concerning text processing, transcripts of the IPUs are processed by the CKIP automatic word segmentation and POS tagging system [8]. Word segmentation results are then integrated with phone boundary information generated by the ILAS phone aligner to derive word boundary information in speech data. The flowchart of our corpus construction procedures is shown in Figure 1. A preliminary version of spoken language resources with multilayer linguistic annotation is obtained in the PRAAT format [5]. Subsequently, the boundaries of paralinguistic sounds are examined manually to eliminate misalignment caused by some of the voiced paralinguistic sounds in the adjacent speech sounds.

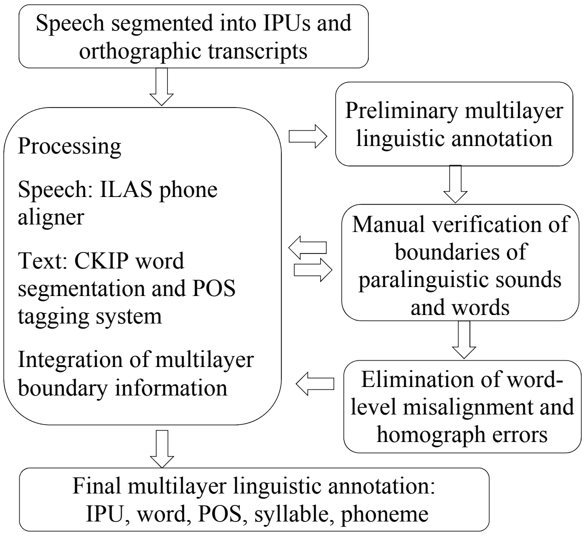

Speech segmented into IPUs and orthographic transcripts
Processing
Speech: ILAS phone aligner
Text: CKIP word segmentation and POS tagging system
Integration of multilayer boundary information
Preliminary multilayer linguistic annotation
Manual verification of boundaries of paralinguistic sounds and words
Elimination of word-level misalignment and homograph errors
Final multilayer linguistic annotation: IPU, word, POS, syllable, phoneme

Figure 1: *Construction of Chinese Spoken Corpora*

After the correction of paralinguistic sound boundaries and a second-round forced alignment, the resulting word boundaries are manually verified by professional labelers, including incorrect word segmentation resulting from unknown words, fragmentary utterances, and disfluencies. Homographs that are very common in Chinese are corrected, if necessary, by referring to the Chinese Spoken Wordlist that contains manually corrected phonetic transcription. This step is essential, as incorrectly converted phonetic transcriptions (from Chinese characters) including tone labels would directly lead to incorrect phone alignment results. After the manual editing process is completed, forced alignment is conducted to accomplish the final multilayer linguistic annotation, as shown in Figure 2. It consists of the levels of IPU, word, POS, syllable, and phoneme. Please note that as a by-product, the annotation schema also contains a Hanyu Pinyin transcription with tone information. Hanyu Pinyin is a phonetic transcription convention in the Romanized alphabet used nationwide in China and in academic papers in Taiwan.

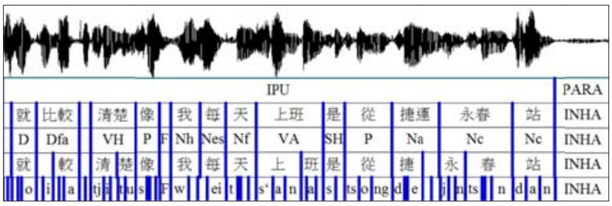

IPU
PARA
就 比較 清楚 像 我 每 天 上班 是 從 捷運 永春 站
INHA
D Dfa VH P F Nh Nes Nf VA SH P Na Nc Nc
INHA
就 較 清 楚 像 我 每 天 上 班 是 從 捷 永 春 站
INHA
o i a tj tu s F w ei t s' a n a s ts o ng d e j n ts n d a n
INHA

Figure 2: *Multilayer linguistic annotation schema*

2.3. ILAS phone aligner

A Chinese syllable has four phonemic positions at most, i.e., CGVX, where C is a consonant onset, G is a glide, V is a vowel, and X is a nasal coda [10] [16]. The phone inventory used in the ILAS phone aligner consists of 22 consonants, /p, pʰ, m, f, t, tʰ, n, l, s, ts, tsʰ, ʂ, tʂ, tʂʰ, ʐ, ɕ, tɕ, tɕʰ, x, k, kʰ, ŋ/, 2 glides /j, w/, and 15 vowels /i, ɨ, u, ɯ, y, a, o, ə, e, ɚ, ai, ei, au, ou, ye/. It was trained with 51 tri-state acoustic models using the HTK toolkits [44], including 39 models for the phone inventory, four for discourse particles, three for fillers of different forms, three for paralinguistic sounds, one for word fragment, and one for foreign words, e.g., English words [23] [19]. The initial acoustic models were trained and fine-tuned using the Sinica Phone-aligned Conversational Speech Database (SPCCSD) with manually labeled phone boundaries. The ILAS phone aligner has reliable performance with a marginal deviation from the human-verified boundaries. In the corpus construction procedures introduced above, please note that as the phone sequences are automatically generated from the citation form of the transcribed words, all phonemes are present in the annotation tier, including those that are actually omitted or extremely reduced. As a minimum length of 15 milliseconds was assigned to each phone model due to the 5 milliseconds of frame shift and three emitting states with no skips of the aligner construction in the Sinica phone aligner, any phone that has a length of 15 milliseconds in our corpus can be considered omitted.

2.4. Word segmentation and POS tagging

The transcripts contain Chinese characters with no blanks in between. While transcribing the contents of the conversations, the transcribers do not segment words by themselves, as word segmentation in Chinese may vary from one morphological theory to another. For this reason, we did not apply any theory-based criteria for word segmentation but adopted the CKIP automatic word segmentation and POS tagging system for processing our transcripts. The CKIP system is trained on a set of words from Chinese dictionaries and a text corpus (the Sinica Balanced Corpus) [8] as a basis, enhanced with morphological rules for word and compound derivation. For processing spoken Chinese data, we slightly modified the ranking of rules and enhanced the dictionary with lexical information from the Sinica Chinese Spoken Wordlist so that the CKIP system is well accommodated to our conversational speech data. Herewith, we would like to express our sincere gratitude to Professor Keh-Jiann Chen for providing the technical aids. Incorrectly segmented words were manually edited during the working phase of word boundary verification. Table 1 lists the categorization of word classes used in our spoken corpus construction in terms of the CKIP POS tagset. For a detailed definition of the POS tagset, please refer to the official website of the CKIP at https://ckip.iis.sinica.edu.tw.

Table 1: *Word class and CKIP POS tagset*

Word class	CKIP POS tags
Adjectives	Nonpredicative adjective (A)
Adverbs	Adverb (D), quantitative adverb (Da), preverbal adverb of degree (Dfa), postverbal adverb of degree (Dfb), sentential adverb (Dk), aspectual adverb (Di)
Conjunctions	Coordinate conjunction (Caa), correlative conjunction (Cbb), conjunction: *dengdeng* (Cab), conjunction: *dehua* (Cba)
Determinatives	Demonstrative determinatives (Nep), quantitative determinatives (Neqa), specific determinatives (Nes), numeral determinatives (Neu), postquantitative determinatives (Neqb)
Foreign words	Foreign words (FW)
Interjections	Interjection (I)
Classifier	Measure (Nf)
Nouns	Common noun (Na), proper noun (Nb), place noun (Nc), localizer (Ncd), time noun (Nd), postposition (Ng), nominalization (Nv)
Particles	Particle (T)
Prepositions	Preposition (P)
Pronouns	Pronoun (Nh)
Sentence	Nominal expression, idioms (S)
Verbs	Active intransitive verb (VA), active pseudotransitive verb (VB), stative intransitive verb (VH), stative pseudotransitive verb (VI), active causative verb (VAC), active transitive verb (VC), active verb with a locative object (VCL), ditransitive verb (VD), active verb with a sentential object (VE), active verb with a verbal object (VF), classificatory verb (VG), stative causative verb (VHC), stative transitive verb (VJ), stative verb with a sentential object (VK), stative verb with a verbal object (VL), 有 *you* (V_2)
DE	Structural particles
SHI	Copula

2.5. Sinica Chinese Spoken Wordlist

The Sinica Chinese Spoken Wordlist (Sinica File No. 24T-1080124) was mainly derived from the transcripts of the adult conversational speech corpora (the TMC), supplemented with a small number of words from other sources. For the step of word segmentation verification, the annotators were instructed to follow the segmentation rules below to modify or correct the results generated from the CKIP system.

(1) Directional complements: 看 不 出來 (*kan bu chulai*)
(2) Proper nouns: 壹週刊 (*yizhoukan*)
(3) Grammatical repetition: 炒作炒作 (*chaozuochaozuo*)
(4) Disfluency repetition: 炒作 炒作 (*chaozuo chaozuo*)
(5) *Dehua* as a word unit: 是 你 的話 (*shi ni dehua*)
(6) *De* as a word unit: 你 的 話 我 不 信 (*ni de hua wo bu xin*)
(7) Numbers as a word unit: 十萬 元 (*shiwan yuan*)

In addition, the phonetic transcription of the words was thoroughly examined, including homographs and tone variants, e.g., 一 (*yi*) and 不 (*bu*). As a result, the Wordlist provides Hanyu Pinyin with tone category labels, POS, frequency and accumulated frequency of ordinary words (405,435 tokens, 16,683 types). Frequency information about discourse-related items is provided in three categories of discourse markers, particles, and fillers (57,697 tokens, 124 types). In addition, statistics in terms of syllable structures and tone categories are included in the Wordlist. Derived from the word information out of the Wordlist, the Sinica Chinese Syllable Structure Frequency List (Sinica File No. 24T-1080125) contains eligible onset-rhyme combinations in Chinese with frequency information. For applications in linguistic research, it can be used for designing stimuli in experiments that are related to the authentic use of spoken Chinese. Both resources have been released for free academic licensing since 2012 by Academia Sinica.

2.6. Query system

Currently, we are developing an online query system that can be used to access the spoken language resources introduced in this paper.

3. Adult conversation corpus

3.1. TMC Corpus settings

The 43-hour Taiwan Mandarin Conversational Corpus (TMC Corpus) consists of 30 free conversations between strangers and 29 topic-specific and 26 Map Task conversations between people acquainted with each other, each an average length of one hour, 20 minutes, and 10 minutes, respectively. For further details regarding the corpus scenario settings and the individual topics of the conversations, please refer to [38]. The TMC Corpus has a balanced design of scenarios and conversation partner familiarity. Ninety-eight female and 72 male speakers aged 16 to 63 years were recorded. Twenty-six speakers took part in all three subcorpus projects. Conversations were recorded in quiet rooms in Academia Sinica by using the SONY TCD-D10 Pro II DAT digital recorder and the Audio-Technica ATM 33a microphone at a sampling rate of 48 kHz, with each speaker on a separate channel. The speech content was orthographically transcribed using traditional Chinese characters. Particles, discourse markers, fillers, word fragments, and paralinguistic sounds that often occur in Chinese conversation are accordingly annotated in the transcripts.

3.2. TMC Corpus statistics

With lexical information from the transcripts and sound files, the working procedures illustrated in Figure 1 were conducted step by step to obtain the final version of the annotated speech corpora with multilayer linguistic information. Please note that all word boundaries were manually verified before the final step of forced alignment. The details of the TMC Corpus are summarized in Table 2. Since 2014, two subsets of the TMC have been released for academic use, the Sinica MCDC8 (Sinica File No. 24T-1031223) and the Sinica Phone-aligned Chinese Conversational Speech Database (SPCCSD, Sinica File No. 24T-1031221). The former consists of eight hours of conversational speech with signal-aligned IPUs and word boundary annotation. The latter contains 3.5 hours of speech data with manually verified phone boundary information. Academic licenses can be issued via the Association for Computational Linguistics and Chinese Language Processing (http://www.aclclp.org.tw/).

Table 2: *TMC Corpus details: Token (type, POS type)*

IPU	81,327
Word	*Lexical words: 397,693 (15,105)*
	1-syllabic words: 224,343 (1,580)
	2-syllabic words: 153,240 (9,705)
	3-syllabic words: 17,322 (2,942)
	Others: 2,788 (878)
	Discourse-related items: 175,318 (2,419)
	Discourse particles: 29,421 (36)
	Discourse markers: 12,164 (16)
	Fillers: 16,721 (34)
POS	Verbs: 98,090 (6,261, 16)
	Adverbs: 80,190 (657,64)
	Nouns: 75,559 (8,210,7)
	Pronouns: 39,453 (50, 1)
	Determinatives: 24,865 (526, 5)
	Preposition: 14,464 (100, 1)
	Conjunctions: 17,950 (94, 4)
	Structural particles DE: 16,342 (5, 1)
	Classifiers: 12,969 (165, 1)
	Particles: 3,802 (22, 1)
	Adjectives: 813 (193, 1)
	Interjection: 8 (4, 1)
	Copula: 13,141 (3, 1)
	Foreign words: 1,470 (473, 1)
Character	594,238 (2,952)
Syllable	Tone-distinctive 1,086
	No tone distinction 403
Phoneme	1,429,518

Current annotation projects conducted for the Sinica MCDC8 include Universal Dependency POS conversion, automatically derived word chunks, discourse units, prosodic units, disyllabic word reduction types and filler types. We warmly welcome colleagues who are interested in collaborative work in related research fields to contact us.

4. Sinica Sociophonetic Corpus

4.1. Corpus settings

The Sinica Sociophonetic Corpus was funded by the National Digital Archives Project. The purpose of this corpus project was to document and archive the contemporary use of spoken

Taiwan Mandarin [41] [36]. Recording was conducted in twelve regions distributed across northern, middle, and southern Taiwan, including Yilan County, Taoyuan County, Hsinchu County, Taichung City, Nantou County, Yunlin County, Chiayi City, Changhua County, Tainan City, Kaohsiung City, Kaoshiung County, and Taipei City. A total of 1,402 interviews mainly with individuals aged 20 to 40 years were recorded in public places, e.g., parks, post offices, or banks, where we assumed we were most likely to find local people. The interviews were recorded by using the Sony Hi-MD MZ-RH1 digital recorder and the Sony ECM MS907 microphone, digitized at a sampling rate of 44.1 kHz with 16-bit quantization. The speech content of the interviewees was orthographically transcribed in traditional Chinese characters with annotations of paralinguistic sounds and pauses. Speech signal alignment was conducted as mentioned above.

Twenty-five questions in three categories were directed to the interviewees, including information about the language use, socioeconomic background, and use of the internet of the interviewees. Concerning language use, dialect exposure was particularly specified in the way the spoken dialects, mainly Southern Min and Hakka, are used within a family, e.g., to parents and siblings. Questions about language ability are concerned with how many languages the interviewees can speak and how good they are. Please note that all dialects were counted as distinctive languages in this context. Concerning socioeconomic background, data on age, gender, salary level, education level and childhood residence were sought. The length of individual interviews ranged from three to eight minutes. All interviews were conducted in Taiwan Mandarin.

4.2. Corpus statistics

Only the speech produced by the interviewees was transcribed and processed. The final statistics are listed in Table 5. In addition to speech processing, the answers to each of the questions were coded into a database containing subjects' socioeconomic information for further sociophonetic studies.

Table 5: *Sociophonetic corpus details*

IPU	124,916
Word	*Lexical words: 284,196 (7,085)*
	1-syllabic words: 133,354 (1,007)
	2-syllabic words: 129,060 (4,218)
	3-syllabic words: 20,348 (1,585)
	Others: 1,434 (275)
	Discourse-related items: 122,634 (718)
	Discourse particles: 28,928 (33)
	Discourse markers: 3,993 (12)
	Fillers: 28,826 (21)
POS	Verbs: 58,894 (2,135, 16)
	Adverbs: 42,750 (367, 6)
	Nouns: 88,146 (4,423, 7)
	Pronouns: 10,020 (33, 1)
	Determinatives: 18,700 (362, 5)
	Preposition: 11,499 (66, 1)
	Conjunctions: 9,817 (64, 4)
	Structural particles DE: 6,655 (7, 1)
	Classifiers: 5,917 (76, 1)
	Particles: 6,324 (19, 1)
	Adjectives: 683 (102, 1)
	Interjections: 3 (2, 1)
	Copula: 10,275 (4, 1)
	Foreign words: 2,579 (235, 1)
Character	458,320 (2,006)
Syllable	Tone-distinctive 929 No tone distinction 375
Phoneme	1,102,753

5. Sinica Child Speech Corpus

5.1. Subjects and corpus settings

The Sinica Child Speech Corpus contains repetitive and narrative speech data produced by seventy-nine preschool children with normal hearing (NH) aged 2;11~6;3 (median 5;0) and forty-five children with hearing impairment (HI) aged 3;3~12;5 (median 5;9). Among the HI children, thirty wore traditional hearing aids (with mild to profound degrees of hearing loss), and fifteen were fitted with a cochlear implant (with severe to profound degrees of hearing loss). The HI children were recorded during their regular AVT session [11] using the video equipment built into the sound-proof classrooms of the Children's Hearing Foundation. Adobe Audition 1.0 was used to convert the video files into 44100 Hz, 16-bit single-channel sound files. To ensure that the HI children had no difficulties hearing or understanding the vocabulary used in the 18 test sentences, the audiologists of the Children's Hearing Foundation conducted an additional word recognition test with each of the HI children shortly before or after our recording sessions. The NH children were recorded either at Academia Sinica in sound-proof studios or in quiet classrooms at their kindergarten using the Sony Hi-MD MZ-RH1 digital recorder and the Sony ECM MS907 microphone. The data were digitized at a sampling rate of 44.1 kHz with 16-bit quantization. All NH and HI were also scored for their overall intelligibility for research on the correlates of spoken language ability and acoustic properties.

5.2. Repetitive speech recording

Eighteen sentences listed in Table 3 that contain all Chinese phonemes with no consideration of their position within the syllable were recorded by a female adult speaker and then played to the children one by one. For each sentence, the children also had visual information on the computer screen illustrating cartoon pictures matching the meaning of the sentences, as shown in Figure 3. Then, they were instructed to repeat the sentences they heard one by one. All recorded speech data were processed following the procedures in Figure 1.

Table 3: *Sentences used in collecting repetitive speech data*

1 我可以看電視 *wo keyi kan dianshi*	"I can watch TV"
2 我們可以看電視 *women keyi kan dianshi*	"We can watch TV"
3 請你把書給我 *qing ni ba shu gei wo*	"Please pass me the book"
4 請你們把書給我 *qing nimen ba shu gei* wo	"Please pass me the book"
5 他喜歡蛋糕 *ta xihuan dangao*	"He likes cakes"
6 他們喜歡跳舞 *tamen xihuan tiaowu*	"They love dancing"
7 三個大人 *san ge daren*	"Three adults"
8 一個小孩 *yi ge xiaohai*	"A child"
9 一隻小雞 *yi zhi xiaoji*	"A chicken"
10 一盒彩色筆 *yi he caiseb*	"A box of coloring pencils"
11 不會來 *bu hui lai*	"Someone won't come"
12 我會寫字 *wo hui xiezi*	"I can write"
13 我會自己吃飯 *wo hui ziji chifan*	"I can eat by myself"
14 黑色的狗很漂亮 *heise de gou hen piaoliang*	"Black dogs are beautiful"
15 喜不喜歡恐龍 *xi bu xihuan konglong*	"Do you like dinosaurs or not"
16 他喜歡黃色 *ta xihuan huangse*	"He likes yellow"
17 他喜歡黃色的蛋糕 *ta xihuan huangse de dangao*	"He likes yellow cakes"
18 不可以淋雨 *bu keyi linyu*	"Don't stay in the rain"

Figure 3: *Pictures for sentences 2, 6, 13, and 18 (Yi-zhen Lin)*

5.3. Storytelling speech recording

For narrative speech data collection, the children were asked to tell two stories, *Little Bear Brings an Apple* and *The Hare and the Tortoise*, assisted with picture cards that were presented to them in a fixed order, as shown in Figure 4.

Figure 4: *The Little Bear Brings an Apple (Xinyi Publisher) and the Hare and the Tortoise (Lishangte Publisher)*

The speech content was orthographically transcribed in traditional Chinese characters with annotations of paralinguistic sounds and pauses [39] [40]. As the story *Little Bear Brings an Apple* is rather simple and not quite suitable for observing the complex syntactic structure patterns of the children for the current project, we only processed the speech data for *The Hare and the Tortoise.* The corpus statistics are summarized in Table 4.

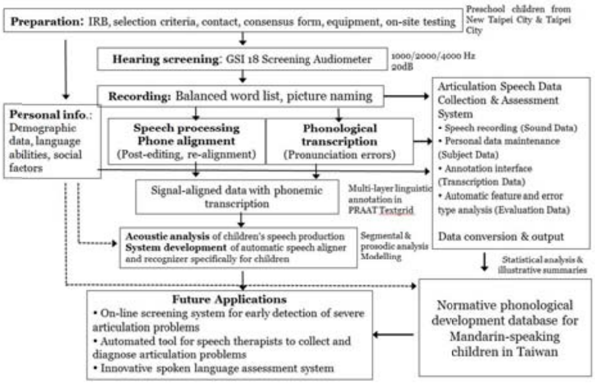
Table 4: *Story-telling speech data details*

		HI	NH
IPU		2,208	2,727
Word	*Lexical words*	*5,193 (503)*	*6,436 (559)*
	1-syllabic words	3,002 (181)	3,863 (205)
	2-syllabic words	2,123 (276)	2,484 (305)
	3-syllabic words	61 (40)	86 (46)
	Others:	7 (6)	3 (3)
	Discourse-related items	*2,778 (42)*	*3,695 (51)*
	Discourse particles	75 (14)	52 (16)
	Discourse markers	21 (4)	53 (8)
	Fillers	56 (8)	214 (11)
POS	Verbs	1,612 (252, 16)	1,857 (287, 16)
	Adverbs	1,038 (71, 6)	1,388 (75, 6)
	Nouns	1,209 (135, 7)	1,254 (148, 7)
	Pronouns	334 (10, 1)	650 (11, 1)
	Determinatives	251 (25, 5)	344 (28, 5)
	Preposition	155 (15, 1)	241 (20, 1)
	Conjunctions	79 (15, 4)	92 (15, 4)
	Structural particles DE	131 (2,1)	107 (2, 1)
	Classifiers	163 (7, 1)	239 (8, 1)
	Particles	170 (10, 1)	183 (10, 1)
	Adjectives	2 (1, 1)	2 (1, 1)
	Interjections	0	0
	Copula	49 (1, 1)	77 (2, 1)
Character		7,467 (378)	9,102 (408)
Syllable	Tone-distinctive	311	349
	No tone distinction	215	236
Phoneme		17,046	21,022

6. Child phonological development corpus

6.1. Child speech collection and assessment project

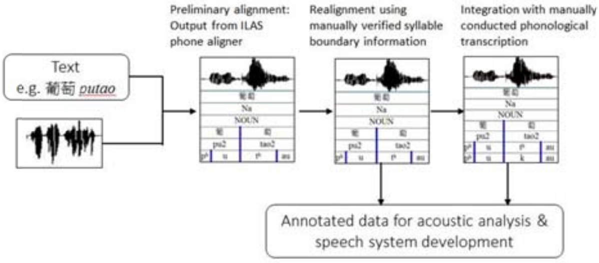
To collect the child phonological development speech data, a commercial system was used with some of our own adaptations to integrate multiple functions of personal data management, sound recording, phonological transcription, and automatic analysis of phonological development patterns as well as to manage large-scale speech data processing. Recording was conducted using the MacBook Air Pro Retina 13.3 and the Sony ECM MS907 microphone. The data were digitized at a sampling rate of 16 kHz. Subject data include information about age, gender, hearing ability, language ability and other demographic data. The online recording platform facilitates the illustration of pictures with the corresponding texts. An interactive working interface is available for phonological transcription of the recorded speech content. For speech data processing, the processes described in Figure 1 will be completed for training an automatic speech recognition system targeted for domain-specific applications. With information about phone boundaries, acoustic features and their correlates with sensory properties will be further analyzed in upcoming research projects. The overview of the project is illustrated in Figure 5.



Preparation: IRB, selection criteria, contact, consensus form, equipment, on-site testing
Preschool children from New Taipei City & Taipei City
Hearing screening: GSI 18 Screening Audiometer
1000/2000/4000 Hz 20dB
Recording: Balanced word list, picture naming
Personal info.: Demographic data, language abilities, social factors
Speech processing Phone alignment (Post-editing, re-alignment)
Phonological transcription (Pronunciation errors)
Articulation Speech Data Collection & Assessment System
• Speech recording (Sound Data)
• Personal data maintenance (Subject Data)
• Annotation interface (Transcription Data)
• Automatic feature and error type analysis (Evaluation Data)
Data conversion & output
Signal-aligned data with phonemic transcription
Multi-layer linguistic annotation in PRAAT Textgrid
Acoustic analysis of children's speech production
System development of automatic speech aligner and recognizer specifically for children
Segmental & prosodic analysis Modelling
Statistical analysis & illustrative summaries
Future Applications
• On-line screening system for early detection of severe articulation problems
• Automated tool for speech therapists to collect and diagnose articulation problems
• Innovative spoken language assessment system
Normative phonological development database for Mandarin-speaking children in Taiwan

Figure 5: *Phonological development corpus project*

It is also planned that the accuracy of segmental (and probably also tonal) pronunciation in terms of the sound inventory and tones of Chinese will be assessed by trained phoneticians or speech therapists. Phone boundary annotation obtained by the ILAS phone aligner will be manually verified at the syllable level. Eventually, the annotation tiers will be integrated with the phonological transcription results, leading to the final version of the phone boundary information (signal alignment) and phonological transcription (actual pronunciation), as shown in Figure 6.



Preliminary alignment: Output from ILAS phone aligner
Realignment using manually verified syllable boundary information
Integration with manually conducted phonological transcription
Text
e.g. 葡萄 putao
Na
NOUN
pu2
tao2
Annotated data for acoustic analysis & speech system development

Figure 6: *Integrated annotation tiers*

6.2. A balanced design of wordlist

The wordlist used for recording consists of 62 disyllabic and 8 trisyllabic words. All onsets and rhymes eligible for composing Chinese syllables occur in both the first and second syllable positions. Except for the neutral tone, all 2x2 combinations of tones in disyllabic words are considered in the wordlist. For composing child-friendly sentences and short discourse content for the future recording of continuous speech, we specifically considered a number of different semantic fields of words that are familiar to children in the wordlist, including animals, food, transportation, body parts, movement, objects, games, locations, and natural phenomena. The wordlist is shown in Table 6 [34] [35].

Table 6: *Sinica Child Balanced Wordlist*

Word	Pinyin	Tones	Word	Pinyin	Tones
母雞	*muji*	3_1	騎馬	*qima*	2_3
蜜蜂	*mifeng*	4_1	走路	*zoulu*	3_4
恐龍	*konglong*	3_2	關燈	*guandeng*	1_1
老鷹	*laoying*	3_1	掃地	*saodi*	3_4
烏龜	*wugui*	1_1	睡覺	*shuijiao*	4_4
兔子	*tuzi*	4_5	買菜	*maicai*	3_4
老虎	*laohu*	3_3	爬山	*pashan*	2_1
螃蟹	*pangxie*	2_4	穿衣服	*chuanyifu*	1_1_2
蜘蛛	*zhizhu*	1_1	電視	*dianshi*	4_4
天鵝	*tiane*	1_2	輪胎	*luntai*	2_1
刺蝟	*ciwei*	4_4	窗戶	*chuanghu*	1_4
醜小鴨	*chouxiaoya*	3_3_1	吸管	*xieguan*	1_3
熱狗	*regou*	4_3	時鐘	*shizhong*	2_1
饅頭	*mantou*	2_2	筷子	*kuaizi*	4_5
蛋糕	*dangao*	4_1	茶杯	*chabei*	2_1
芒果	*mangguo*	2_3	皮鞋	*pixie*	2_2
果汁	*guozhi*	3_1	玩具	*wanju*	2_4
牛奶	*niunai*	2_3	鈕扣	*niukou*	3_4
草莓	*caomei*	3_2	盤子	*panzi*	2_5
葡萄	*putao*	2_2	彩色筆	*caisebi*	3_4_4
牛排	*niupai*	2_2	溫度計	*wenduji*	1_4_4
蘋果	*pingguo*	2_3	足球	*zuqiu*	2_2
壽司	*shousi*	4_1	拼圖	*pintu*	1_2
甜甜圈	*tiantianquan*	2_2_1	積木	*jimu*	1_4
汽車	*qiche*	4_1	大富翁	*dafuweng*	4_4_1
飛機	*feiji*	1_1	吹泡泡	*chuipaopao*	1_4_4
火車	*huoche*	3_1	學校	*xuexiao*	2_4
耳朵	*erduo*	3_1	廚房	*chufang*	2_2
牙齒	*yachi*	2_3	客廳	*keting*	4_1
嘴巴	*zuiba*	3_1	花園	*huayuan*	1_2
說話	*shuohua*	1_4	噴水池	*penshuichi*	1_3_2
寫字	*xiezi*	3_4	白雲	*baiyun*	2_2
吃飯	*chifan*	1_4	月亮	*yueliang*	4_4
淋雨	*linyu*	2_3	斷崖	*duanyai*	4_2
游泳	*youyong*	2_3	生日	*shengri*	1_4

6.3. Current project progress

The project is still ongoing. To date, data for 747 preschool children and 141 elementary school children were recorded as listed in Table 7. Some of the data are being annotated and analyzed in terms of the tone contour development pattern, funded by the Ministry of Science and Technology in Taiwan.

Table 7: *Current status (by September 2019)*

	Kindergarten				Elementary school		
Age group	3~4	4~5	5~6	6~7	G1	G2	G3
Male	36	114	129	101	20	29	19
Female	50	122	108	87	21	30	22

7. Acknowledgements

The author is sincerely grateful to the financial support from the Institute of Linguistics, Academia Sinica (ILAS), the National Digital Archives Project (NDAP), the National Science Council (NSC), the Ministry of Science and Technology (MOST), and the Children's Hearing Foundation (CHF) as well as the contributions of assistants, students, programmers and the staff of the participating schools and kindergartens who have been working very hard and have been dedicated to their specific tasks in the last two decades in constructing the corpora presented in this paper.

8. References

[1] Anderson, A. H., Bader, M., Bard, E. G., et al. (1991). The HCRC map task corpus. Language and speech, 34(4), 351-366.

[2] Artstein, R. & Poesio, M. (2008). Inter-coder agreement for computational linguistics. Computational Linguistics, 34(4), 555-596.

[3] Bally, C. (1952). Le langage et la vie. 3rd ed. Genève: Droz. [Bally, Charles. 1913. Ferdinand de Saussure et l'état actuel des études linguistiques. Genève: Atar. (Reprinted in Bally 1952: 147–160.) [Inaugural lecture for the chair of general linguistics, 27 Oct. 1913.]

[4] Bigi, B. & Hirst, D. (2012). Speech Phonetization Alignment and Syllabification (SPPAS): a tool for automatic analysis of speech prosody. In Proceedings of Speech Prosody (pp. 19-22). Shanghai.

[5] Boersma, P. & Weenink, D. (2013). Praat: Doing phonetics by computer. Retrieved from http://www.praat.org/ (last access May 2013).

[6] Brown, G. D. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. Behavior Research Methods, Instruments, & Computers, 16(6), 502-532.

[7] Chao, Y.-R. (1968). A grammar of spoken Chinese. University of California Press.

[8] Chen, K.-J., Huang, C.-R. Chang, L.-P. & Hsu, H.-L. (1996). SINICA CORPUS: Design methodology for balanced corpora. In Proceedings of the Eleventh Pacific Asia Conference on Language, Information, and Computation (pp. 167-176). Seoul.

[9] Cohen, J. (1960). A Coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46.

[10] Duanmu, S. (2000). The phonology of standard Chinese. New York: Oxford University Press.

[11] Dornan, D., Hickson, L., Murdoch, B. & Houston, T. (2007). Outcomes of an auditory-verbal program for children with hearing loss: A comparative study with a matched group of children with normal hearing. The Volta Review, 107, 37-54.

[12] Effers, B., Van Bael, C. & Strik, H. (2005). Algorithm for Dynamic Alignment of Phonetic Transcriptions. Technical Report. Department of Language and Speech. Radboud University Nijmegen. The Netherlands.

[13] Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. Reprinted in F.R. Palmer (ed.). 1968. Selected papers of J.R. Firth 1952-1959. London: Longman.

[14] Godfrey, J. J., Holliman, E. C. & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (pp. 517-520). San Francisco.

[15] Greenberg, S., Hollenback, J. & Ellis, D. (1996). Insights into Spoken Language Gleaned from Phonetic Transcription of the Switchboard Corpus. In Proceedings of the International Conference on Spoken Language Processing (pp. S24-27). Philadelphia.

[16] Ho, D.-a. (1996). Some Concepts and Methodology of Phonology. Da-An Press. (in Chinese)

[17] Jurafsky, D. & Martin, J. (2008). Speech and Language Processing: An Introduction to Natural Language Processing,

Computational Linguistics, and Speech Recognition (2nd ed.). Pearson Education India.

[18] Kohler, K. J. (1996). Labelled data bank of spoken standard German: the Kiel corpus of read/spontaneous speech. In Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96 (Vol. 3, pp. 1938-1941). IEEE.

[19] Kondrak, G. (2003). Phonetic alignment and similarity. Computers and the Humanities, 37, 273-291.

[20] Levelt, W. J. M. (1993). Speaking: From intention to articulation. Vol. 1. MIT Press.

[21] Li, A., Zheng, F., Byrne, W., Fung, P., Kamm, T., Liu, Y., Song, Z., Ruhi, U., Venkataramani, V. & Chen, X. (2000). CASS: A phonetically transcribed corpus of Mandarin spontaneous speech. In Sixth International Conference on Spoken Language Processing.

[22] Liu, Y. & Fung, P. (2004). State-dependent Phonetic Tied Mixtures with Pronunciation Modeling for Spontaneous Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing, 12(4), 351-364.

[23] Liu, Y.-F., Tseng, S.-C. & Jang, J.-S. R. (2014). Phone boundary annotation in conversational speech. In Proceedings of the International Conference on Language Resources and Evaluation (pp. 848-853). Reykjiavik.

[24] Maekawa, K., (2003). Corpus of Spontaneous Japanese: Its design and evaluation. In Proceedings of ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (pp. 7-12). Tokyo.

[25] Oostdjik, N., (2000). The Speech Dutch Corpus. Overview and First Evaluation. In Proceedings of the International Conference on Language Resources and Evaluation (pp. 887-894). Athens.

[26] Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In Knowledge of variation. In J. L. Bybee and P. J. Hopper (Eds.), Frequency and the Emergence of Linguistic Structure (pp. 137-158). John Benjamins Publishing Company.

[27] Pitt, M. A., Johnson, K., Hume, E., Kiesling, S. & Raymond, W. (2005). The Buckeye Corpus of Conversational Speech: Labeling Conventions and a Test of Transcriber Reliability. Speech Communication, 45, 89-95, 2005.

[28] Pluymaeker, M., Ernestus, M. & Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch, Journal of the Acoustical Society of America, 118, 2561-2569.

[29] Schuppler, B., Ernestus, M., Scharenborg, O. & Boves, L. (2011). Acoustic Reduction in Conversational Dutch: A Quantitative Analysis based on Automatically Generated Segmental Transcriptions. Journal of Phonetics, 39, 96-109.

[30] Son, R. V., Binnenpoorte, D., van den Heuvel, H. & Pols, L. (2001). The IFA corpus: A phonemically segmented Dutch 'open source' speech database. In Proceedings of Eurospeech (pp. 2051–2054). Aalborg.

[31] Strik, H. & Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature, Speech Communication, 29, 225-246.

[32] Svartvik, J. & Quirk, R. (1980). A Corpus of English Conversation. Lund, Sweden. Gleerup.

[33] Tsai, M.-Y., Chou, F.-C. & Lee, L.-S. (2007). Pronunciation Modeling with Reduced Confusion for Mandarin Chinese Using a Three-Stage Framework. IEEE Transactions on Audio, Speech, and Language Processing, 15(2), 661-675.

[34] Tseng, S.-C. (2018). A Corpus-based Computational Approach to Collecting and Analyzing Child Speech. The Hong Kong Speech and Hearing Symposium (pp. 21). Hong Kong.

[35] Tseng, S.-C. & Liu, Y.-F. (2017). Establishment of normative phonological development data for Mandarin-speaking children by applying computational linguistics techniques. Presented in the annual meeting of the Speech-Language-Hearing Association of Taiwan (pp. 6). Kaohsiung.

[36] Tseng, S.-C. (2016). /kwo/ and /y/ in Taiwan Mandarin: Social Factors and Phonetic Variation. Language and Linguistics, 17(3), 383-405.

[37] Tseng, S.-C. (2014). Chinese Disyllabic Words in Conversation. Chinese Language and Discourse, 5(2), 231-251.

[38] Tseng, S.-C. (2013). Lexical Coverage in Taiwan Mandarin Conversation. International Journal of Computational Linguistics and Chinese Language Processing, 18(1), 1-18.

[39] Tseng, S.-C. (2011). Speech production of Mandarin-speaking children with hearing impairment and normal hearing. In the Proceedings of the 17th International Congress of Phonetic Sciences (pp. 2030-2033). Hong Kong.

[40] Tseng, S.-C., Kuei, K. & Tsou, P.-C. (2011). Acoustic characteristics of vowels and plosives/affricates of Mandarin-speaking hearing-impaired children. Clinical Linguistics & Phonetics, 25(9), 784-803.

[41] Tseng, S.-C. (2008). Archiving contemporary Taiwan Mandarin speech. Symposium on IT applications and exchange (pp. 269-277). Zhangjiajie. (in Chinese)

[42] Wang, H. M., Chen, B., Kuo, J. W. & Cheng, S. S. (2005). MATBN: A Mandarin Chinese broadcast news corpus. In International Journal of Computational Linguistics & Chinese Language Processing, Volume 10, Number 2, June 2005: Special Issue on Annotated Speech Corpora (pp. 219-236).

[43] Wester, M., Kessens, J. M., Cucchiarini, C. & Strik, H. (2001). Obtaining Phonetic Transcriptions: A Comparison between Expert Listeners and a Continuous Speech Recognizer. Language and Speech, 44, 377-403.

[44] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., et al. (2006). Technical Report: The HTK book (version 3.4), Cambridge University, Engineering Department.