Computational Linguistics and Chinese Language Processing
Vol. 18, No. 1, March 2013, pp. 1-18

Lexical Coverage in Taiwan Mandarin Conversation

Shu-Chuan Tseng*

Abstract

Information about the lexical capacity of the speakers of a specific language is indispensible for empirical and experimental studies on the human behavior of using speech as a communicative means. Unlike the increasing number of gigantic text- or web-based corpora that have been developed in recent decades, publicly distributed spoken resources, espcially conversations, are few in number. This article studies the lexical coverage of a corpus of Taiwan Mandarin conversations recorded in three speaking scenarios. A wordlist based on this corpus has been prepared and provides information about frequency counts of words and parts of speech processed by an automatic system. Manual post-editing of the results was performed to ensure the usability and reliability of the wordlist. Syllable information was derived by automatically converting the Chinese characters to a conventional romanization scheme, followed by manual correction of conversion errors and disambiguiation of homographs. As a result, the wordlist contains 405,435 ordinary words and 57,696 instances of discourse particles, markers, fillers, and feedback words. Lexical coverage in Taiwan Mandarin conversation is revealed and is compared with a balanced corpus of texts in terms of words, syllables, and word categories.

Keywords: Taiwan Mandarin, Conversation, Frequency Counts, Lexical Coverage, Discourse Items.

1. Introduction

Exchange and communication of thoughts are mainly performed by producing and perceiving/interpreting words, whether in text or speech. In spite of philosophical debates on the concept of words, it is more or less accepted by most of the disciplines working with languages that one of the possibilities of exploring the lexical capacity of the users of a specific language is to examine the distribution of words collected in a large-scale balanced corpus. Different from the lexical entries listed in a dictionary, corpus data provide

* Institute of Linguistics, Academia Sinica, Taipei, Taiwan
E-mail: tsengsc@gate.sinica.edu.tw

information about lexical knowledge of language users that resembles their experiences and abilities in a realistic context. Of this information, word frequency counts are simple and primitive information. Nevertheless, they are directly associated with the lexical capacity of language users in a given scenario. Word frequency is one of the most essential kinds of information when implementing language-related technology tools and systems. Once a reliable word list is available, different computational models can be developed or applied to examine the role lexical knowledge plays in using a language (Baayan, 2001). For pedagogical purposes, word counts based on real corpus data will help prepare authentic learning materials for first and second language learners (Xiao *et al*., 2009; Knowles, 1990; McCarthy, 1999). For research purposes, empirical information about lexical capacity is indispensible for constructing stimuli and testing hypotheses for word- or phonology-related psycholinguistic experiments (Wepman & Lozar, 1973). In each kind of application using the word distribution information mentioned above, it is important that the sources we obtain the information from should resemble the word distribution of tokens and types as authentic language input available to the language users.

Nearly a century ago, Thorndike (1921) listed the 10,000 most widely used English words based on a 4.6-million-word corpus consisting of 41 different sources, which included children's literature, the Bible, classics, elementary school textbooks, and newspapers. The later version extended the list to 30,000 words (Thorndike & Lorge, 1944). The main purpose of these earliest wordlists was to provide word information for teaching English. Nowadays, taking advantage of the latest technology, the amount and scale of textual corpora being collected via digital resources in recent decades have become enormous. The British National Corpus (BNC) contains 100 million English words. Within the corpus data, 90% were based on written texts (Leech *et al*., 2001). The first released version of the American National Corpus (ANC) contained 11.5 million English words, 70% of which were written texts (Reppen & Ide, 2004). Both the BNC and the ANC are balanced corpora. They consist of texts collected from different producers and genres, also including transcripts of spoken language. Purely textual corpora, such as the English Gigaword and the Chinese Gigaword, distributed by the Linguistic Data Consortium (LDC), are mostly collections of newspaper articles, reflecting a specific kind of language user behavior. Nevertheless, to reflect the lexical capacity of language users in natural speech communication, we need a corpus of "naturally produced" conversations with different sociolinguistic designs of speaker relationships and different conversation types. Compared with textual corpora, however, it is considerably more difficult to obtain this kind of corpora.

Collecting and processing speech data cannot be accomplished automatically. The cost of preparing spoken corpora is high, especially when dealing with natural conversations. The types of spoken corpora vary to a large degree, ranging from reading a list of words/texts,

telling a story, executing a task, to free conversation. To take English as an example, a number of conversational corpora have been collected for educational, clinical, or experimental studies of spoken word distribution (French *et al*., 1930; Howes, 1964; Howes, 1966). They have attracted intensive attention, because they provide the most realistic materials to study how people converse to exchange thoughts and perform communication. During the last twenty years, the scale and the application of spoken corpora have been enormously extended. Svartvik and Quirk (1980) published a corpus of English conversation, later known as the London-Lund Corpus of English Conversation. A word frequency count of 190,000 words from the corpus was published four years later (Brown, 1984). Later, a part of the BNC also contained conversations, with a focus on a balanced socio-geographic sampling of speakers of English (Crowdy, 1993).

With the growing number of spoken corpora being or having been processed, the technology and the concept of how to prepare spoken corpora has also been changed accordingly due to the extensive application possibilities and the available software (Gibbon *et al*., 1997). Newly developed spoken corpora, for instance, transcribed with annotation schemes marking targeted linguistic phenomena, time-aligned with speech signals at different linguistic levels, automatically processed for word segmentation and parts of speech tagging on the transcripts, etc., have brought new horizons of how spoken corpora can be used for academic and educational purposes.

2. Taiwan Mandarin Spoken Wordlist

This paper studies the lexical coverage of a Taiwan Mandarin conversational corpus based on the derived *Taiwan Mandarin Spoken Wordlist* and compares it with the Sinica Corpus (Chen & Huang, 1996), which is currently the largest POS-tagged text corpus of Taiwan Mandarin. This section gives an introduction to how the conversational corpus has been collected and processed and how the wordlist has been prepared.

2.1 Taiwan Mandarin Conversational Corpus

The Taiwan Mandarin Conversational Corpus (the TMC Corpus, hereafter) is composed of three sub-corpora of Taiwan Mandarin conversations, which have been processed at the Institute of Linguistics, Academia Sinica (Tseng, 2004). The Mandarin Conversational Dialogue Corpus (the MCDC) is a collection of 30 free conversations between speakers who were meeting for the first time (37 females and 23 males, with ages between 16 and 45). The project was executed in 2001. One year later, 30 speakers from the MCDC speakers were recruited again to record conversations with a person they knew well for the next two corpus collection projects. As a result, 33 female and 27 male speakers whose age ranged from 14 to 63 participated in the project. The Mandarin Topic-oriented Conversation Corpus (the MTCC)

is a collection of topic-specific conversations on selected news or events that took place in the year of 2001. The Mandarin Map Task Corpus (the MMTC) is a collection of task-oriented dialogues, basically following the Map Task design (Anderson *et al*., 1991). Different from the MTCC and the MMTC, the free conversations in the MCDC were more formal, as the conversation partners were strangers. The final version of the TMC Corpus consists of 85 conversations, approximately 42 hours of speech recording. Five conversations were not included in the TMC Corpus because the participants spoke Taiwan Southern Min instead of Taiwan Mandarin to their conversation partners most of the time in their conversations. General information about the corpora is summarized in Table 1.

Table 1. Corpus Description of the TMC Corpus.

Sub-Corpus	No. of Speakers	Length per conversation	Corpus Scenario	Conversation partners
MCDC	60 (37F, 23M)	1 hour	Free conversation	Strangers
MTCC	58 (33F, 25M)	20 minutes	Topic-oriented Conversation	Friends/ relatives
MMTC	52 (28F, 24M)	7 minutes	Map task dialogue	Friends/ relatives

From the viewpoint of speaker relationship, the TMC Corpus contains conversations between strangers and conversations between people who are familiar with each other. From the viewpoint of the speaking situation, the TMC Corpus includes three different scenarios: free conversations, topic-specific conversations, and task-oriented conversations. That is, the TMC Corpus provides speech data of a variety of speaker groups communicating in different speaking styles and situations.

2.2 Corpus Transcription

The speech content of the 85 conversations was orthographically transcribed and carefully cross-checked. Words were transcribed in traditional Chinese characters. Pauses and paralinguistic sounds, such as inhalation, coughing, and laughter, were indicated in the transcripts. Items that are often used in spoken discourse, such as discourse particles, discourse markers, fillers, and feedback words, were transcribed with capital letters for two reasons. On the one hand, we wanted to distinguish these items from ordinary words due to their pragmatic function in conversation. On the other hand, it is not always possible to find the correct, or widely accepted, characters to transcribe these groups of items. For example, well-conventionalized characters are available in the writing notion for most of the discourse particles (Chao 1965) originating from Mandarin Chinese, such as: **A** 啊, **AI YA** 哎呀, **AI YOU** 唉呦, **BA** 吧, **E/EP** 呃, **EN** 嗯, **HAI** 嗨, **HE** 呵, **HEI** 嘿, **HWA** 嘩, **LA** 啦, **LIE/LEI** 咧, **LO** 囉, **MA** 嘛,

NOU/NO喏, **O**喔/噢/哦, **OU**噢, **WA**哇, **WA SAI**哇塞, **YE**耶, **YI** 咦, and **YOU**呦. Nevertheless, some of the very common particles in contemporary Taiwan Mandarin conversation, such as **EIN**, **HAN**, **HEIN**, **HO**, **HYO**, and **HAIN,** originate from Taiwanese Southern Min - a major dialect spoken in Taiwan. For these particles, no widely acceptable characters are available to transcribe them. Capital letters signifying the way of pronunciation were used to transcribe discourse particles of this kind. Different from discourse particles, discourse markers noted in our transcribing system are originally lexical items, i.e. regular words with a matching character in the writing system. When their original semantic meaning is lost and their use becomes essentially pragmatic in conversation, however, they are regarded as a kind of discourse markers. Their function is similar to that of the discourse markers that are generally defined, e.g. *well*, *but*, and *ok* (Schiffrin, 1988), marking emerging structure of conversation. In principle, they are used for a speaker to keep the floor or to stall more time to think of what to say next. Among the discourse markers annotated in the TMC Corpus, **NA** is the most frequently used marker. Originally, 那 (**NA)** was a demonstrative determiner, meaning "that". As a discourse marker, however, it sometimes appears before a proper noun, which is grammatically incorrect in the case of a determiner. This example illustrates the difference between 那 (**NA)** as a determiner and as a discourse marker. As a result, we noted discourse markers of this specific group, including **NA**, **NE**, **NA GE**, **NE GE**, **NEI GE**, **SHEN ME**, and **ZHE GE**.

The third type, fillers and feedback words, themselves do not involve any concrete semantic meaning. Fillers function as discourse markers in a similar way, indicating hesitations in speech flow (Shriberg, 1994). Feedback words are used as a response signal to the conversation partner. Different foci on spoken discourse may lead to diversified terminologies and systems of lexical items, for instance, Chao (1965) may regard some of the fillers and feedback words as interjections, carrying specific intonation contours. Nevertheless, in the TMC Corpus, our preliminary goal was to develop a coherent transcription convention for conversation. Basically, we transcribed them according to their syllable structure, because the surface forms of fillers and feedback words are systematically similar. Prosodic realization may add affined pragmatic interpretations to fillers and feedback words. Nevertheless, in the transcription system, we do not make further distinctions. There are four different sub-groups of fillers and feedback words: Zero onset + Schwa + dental nasal coda (**UHN**, **UHNN**, **UHNHN**), zero onset + Schwa + bilabial nasal coda (**UHM**, **UHMM**, **UHMHM**), dental nasal onset + Schwa + dental nasal coda (**NHN**, **NHNN**, **NHNHN**), and bilabial nasal onset + Schwa + bilabial nasal coda (**MHM**, **MHMM**, **MHMHM**, **MHMHMHM**, **MHMHMHMHM**). When they are produced with more than one syllable, each syllable is presented by a repeated **H**. A repeated nasal coda indicates a prolongation of the coda.

Foreign words, such as English or Japanese, are either written in their original writing convention or the equivalent romanization. Speech stretches containing pronunciation variants and code switching are transcribed in the way that the meaning of the speech content is written in Taiwan Mandarin writing convention.

2.3 Time-aligned Transcripts in PRAAT

The orthographic transcription of the corpus is presented in PRAAT with two tiers (Boersma & Weenink, 2012). The first tier gives information about the speaker identity and the sequence number of the speaker's turn in a coded way, and the transcription of the speech content is presented on the second tier. The boundaries of all speaker turns are time-aligned with the speech signal. **Figure 1** is an extract from the MCDC sub-corpus.

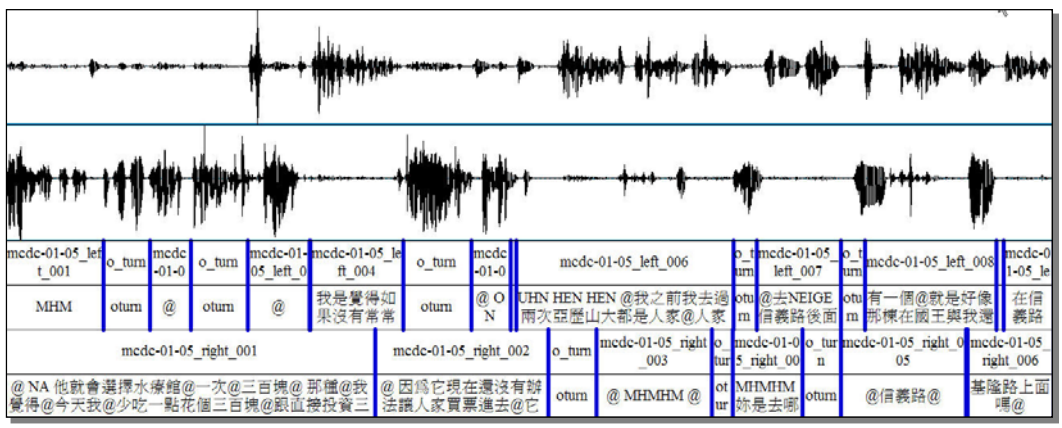

Figure 1. Time-aligned transcription.

2.4 Word Segmentation and POS Tagging

Word boundaries in the Chinese texts are not marked by blanks. In order to prepare the wordlist of the TMC Corpus, we applied the CKIP word segmentation and POS tagging system to automatically process the transcripts (Chen & Huang, 1996). The POS tagset developed by the CKIP team is listed in **Table 2** (CKIP, 1998). Slightly modifying the tagset, we added nominal expressions and idioms to the category S, because they act as independent sentences in conversation from both syntactic and pragmatic points of view and they should not be regarded as any one of the other POS categories. With regard to the input format of the system, the original design of the CKIP system was sentences. For processing the TMC Corpus, the content of each speaker turn was used as individual input to run the CKIP system. As the majority of the corpus data are long speaker turns of more than one sentence, there may arise difficulties in word segmentation and POS tagging. In this regard, manual post-editing would be necessary.

Table 2. The CKIP POS Tagset.

Word category	CKIP POS Tagging system
Adjectives	Non-predicative adjective (A)
Adverbs	Adverb (D), quantitative adverb (Da), pre-verbal adverb of degree (Dfa), post-verbal adverb of degree (Dfb), sentential adverb (Dk), aspectual adverb (Di)
Conjunctions	Coordinate conjunction (Caa), correlative conjunction (Cbb), conjunction:*deng3deng3* (Cab), conjunction:*de5hua4* (Cba)
Determinatives	Demonstrative determinatives (Nep), quantitative determinatives (Neqa), specific determinatives (Nes), numeral determinatives (Neu), post-quantitative determinatives (Neqb)
Foreign words	Foreign words (FW)
Interjections	Interjection (I)
Nouns	Measure (Nf), 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 pseudo-transitive verb (VB), stative intransitive verb (VH), stative pseudo-transitive 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), you3 (V_2)
DE	Structural particles: *de5, zhi1, de2, di4*
SHI	Copula: *shi4*

2.5 Manual Post-editing of Word Segmentation and Homograph Errors

The CKIP word segmentation system was originally trained on written texts. Therefore, incomplete, ungrammatical sentences and peculiar constructions in conversation, which normally do not occur in written texts, could result in errors of the automatic word segmentation and POS tagging system. Segmentation errors, including errors of proper nouns, idioms, constructions with numbers, and directional complements, were manually corrected. In the process of word segmentation and POS tagging, we also need to cope with the occurrences of disfluencies in conversation (Shriberg, 1994). According to the content and the

prosodic realization, a disfluent repetition of words was manually separated (e.g. da uhn da de jiqi, *big uhn big machine*), whereas a grammatical reduplicative phrase was transcribed as one unit (e.g. dadade chengzan ta, *a big compliment to him*).

To obtain information about syllables, all Chinese characters transcribed were automatically converted into Hanyu Pinyin, a romanization convention for Chinese used worldwide. In the system of Hanyu Pinyin, tone information is included with each syllable, which is indicated by 1, 2, 3, 4, and 5, representing Tone 1, Tone 2, Tone 3, Tone 4, and the neutral Tone. Furthermore, because of the large number of homographs in Chinese, post-editing was performed to manually correct errors resulting from the automatic conversion. Ambiguous homographs, which occur very frequently in spoken language, were specified based on the neighboring context. For instance, the word "*one*" (一, yi1) is pronounced with Tone 1 in isolation, but with Tone 2, when followed by Tone 4 and the neutral tone. When followed by Tone 1, Tone 2, and Tone 3, 一 is pronounced with Tone 4. The final version of the automatically segmented and POS tagged words, as well as the manually checked syllables, was used to prepare the wordlist. As a result, the *Taiwan Mandarin Spoken Wordlist*[1] contains 405,435 regular word tokens, equivalent to 16,683 word types and 607,008 syllable tokens. There are 57,696 tokens of discourse particles, discourse markers, fillers, and feedback words.

3. Lexical Coverage in Conversational and Text Corpus

Given a body of language data, no matter in the form of text or speech, lexical coverage revealed from the data varies according to producer- and genre-related factors. Each individual collection of a corpus is only representative of the specific producer group under a given condition of language production. The Sinica Corpus is a balanced corpus of texts containing different genres. In the design of the TMC Corpus, we have attempted to cover varieties of formal and informal speaking situations by the arrangement of conversation partners (strangers vs. familiar persons) and different speaking scenarios by the arrangements of tasks (free conversation, map task, and topic-specific). It is clear that the TMC Corpus and the Sinica Corpus are not directly and completely comparable in terms of producers and genres. Nevertheless, the TMC Corpus and the Sinica Corpus were compiled by adopting the same word segmentation and POS tagging system, and they are currently the largest conversational and textual corpora available for Taiwan Mandarin. For this reason, when we examine the lexical coverage of the TMC Corpus, the Sinica Corpus will be compared to explore the similarities and differences among words produced in the form of conversation and text. Wordlists derived from these two corpora were used, the *Taiwan Mandarin Spoken*

[1] The *Taiwan Mandarin Spoken Wordlist* has been publicly distributed and can be freely downloaded from the website http://mmc.sinica.edu.tw/resources_e_01.htm.

Wordlist and the *Word List with Accumulated Word Frequency in Sinica Corpus 3.0* (CKIP, 1998). In order to collect information about syllables as well, we ran the same automatic conversion program to the *Word List with Accumulated Word Frequency in Sinica Corpus 3.0.* The results, however, were not manually checked, as we did for the TMC Corpus with the homograph errors.

Table 3. Conversational and Text Corpus.

Corpus	Word tokens	Word types	Syllable tokens	Syllable types with tones	Syllable types without tones
TMC Corpus	405,435	16,683	607,008	1,076	390
Sinica Corpus	4,767,048	55,301	7,515,036	1,120	392

For the current study, we have cleaned up errors we found in the wordlist of the Sinica Corpus, so the statistics summarized in **Table 3** may be slightly different from the official ones published by the CKIP team. As one can see, the Sinica Corpus is about ten times bigger than the TMC Corpus.

3.1 Word coverage

Corpus coverage of different vocabulary sizes in both corpora is listed in **Table 4**. The top 2000 word types in the TMC Corpus make up about 90% of the overall word tokens, whereas they only account for 70% of word tokens in the Sinica Corpus. McCarthy (1999: 236) has made a comparable proposal that "… a round-figure pedagogical target of the first 2000 words in order of frequency will safely cover the everyday core with some margin for error." Counting homographs with different POS categories as distinct word types, 1,117 among the top 2000 word types occur in both corpora, including nouns, verbs, adverbs, conjunctions, determinatives, prepositions, pronouns, non-predicative adjectives, particles, the structural particle DE, and the copula SHI. These 1,117 word types shared in the top 2000 list of both corpora eventually account for 81% of the TMC corpus coverage and 58% of the Sinica corpus coverage. A selection of these 1,117 word types, the (approximately) top 100 words in both corpora, is listed in **Appendix A**. They may be regarded as the core vocabulary that is required for operable communication in the form of conversation and text. For educational purposes, this core vocabulary may be the target words for teaching praxis and materials to focus on (Xiao *et al.*, 2009; Tao, 2009).

Table 4. Vocabulary Size and Corpus Coverage.

Vocabulary size	TMC corpus	Tokens	Tokens per type	Vocabulary size	Sinica corpus	Tokens	Tokens per type
1,000	84.43%	342,306	342	1,000	59.78%	2,935,763	2,936
2,000	89.87%	364,364	182	2,000	68.69%	3,373,419	1,687
3,000	92.53%	375,134	125	3,000	73.53%	3,610,951	1,204
4,000	94.20%	381,919	95	4,000	76.77%	3,770,449	943
5,000	95.34%	386,559	77	5,000	79.17%	3,887,992	778
6,000	96.21%	390,081	65	6,000	81.03%	3,979,351	663
7,000	96.95%	393,058	56	7,000	82.55%	4,054,042	579
8,000	97.44%	395,058	49	8,000	83.82%	4,116,688	515
9,000	97.94%	397,058	44	9,000	84.92%	4,170,368	463
10,000	98.35%	398,752	40	10,000	85.87%	4,217,103	422
	100%	**405,435**	**24**		**100%**	**4,767,048**	**86**

The word distribution in both corpora is presented in terms of the accumulative frequency in **Figure 2**. To achieve a 90% of corpus coverage, the first 15,000 frequency-ranked word types in the Sinica Corpus and the first 2,000 ones in the TMC Corpus are required. Calculating the proportions of these word types in their corpus share, 27% of the observed word types in the Sinica Corpus and 12% in the TMC Corpus would account for the majority of the lexical coverage of each corpus. This may suggest that these two different vocabulary sets are required for fluent communication in the form of text and conversation. The size of word types differs largely in both corpora, i.e. 15,000 versus 2,000. Nevertheless, if we view the number of characters involved in the two vocabulary sets, there are 2,964 different characters in the case of the Sinica Corpus and 1,065 in the TMC Corpus. A Chinese character is normally also a morpheme in Mandarin Chinese and is equivalent to a tone-specified syllable. The large number of homographs in Chinese leads to asymmetry between the number of tone-specified syllables from the phonological point of view and the number of characters from the orthographic point of view. The vocabulary sets required for a fluent communication above are equivalent to 1,065 tone-specified syllables for text (1,120 for the Sinica Corpus in total), and 654 for conversation (1,076 for the TMC Corpus in total).

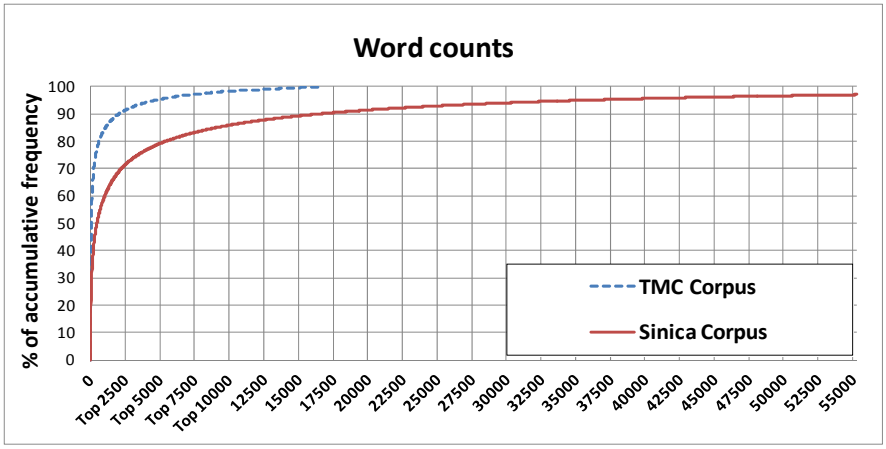

Figure 2. Word distribution.

3.2 Syllable Coverage

In the TMC Corpus, 1,076 different tone-specified syllable types were produced. In the Sinica Corpus, it was 1,120. Apparently, there is no clear difference between the text and conversation corpora in this regard, as shown in **Figure 3**. Similarly, to account for 90% of the corpus coverage, 300 tone-specified syllable types are required in the TMC Corpus and 400 are required in the Sinica Corpus. Moreover, if we disregard tone information, the number of syllable structures is 390 in the TMC Corpus, and 392 in the Sinica Corpus. This is almost the same in both corpora. Among them, 385 syllable structures were found in both corpora and the other 15 syllable structures appeared in only one of the corpora. The figures of syllables in both wordlists suggest that the capacity of phonologically different syllables (with or without considerations of lexical tones) in Taiwan Mandarin used in text and conversation is of similar size. Nevertheless, the number of tone-specified syllables does not equal the number of characters, or morphemes in Mandarin, as we mentioned earlier. For use in the form of text or conversation, the discrepancy is noticable, as the vocabulary sets required for fluent communication differ significantly: 1,065 for text and 654 for conversation.

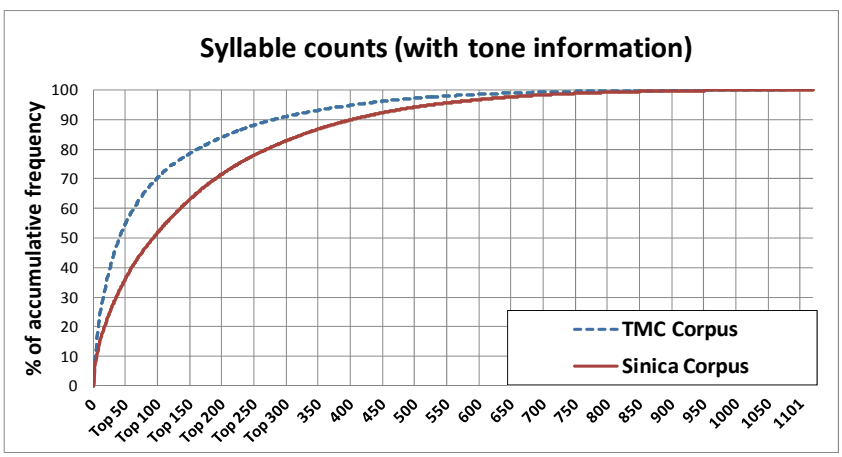

Figure 3. Syllable distribution.

3.3 Distribution of Word Category

The proportions of the 14 categories of the CKIP POS tags in both corpora are summarized in **Table 5**. The occurrences of nouns and verbs in the Sinica Corpus make up nearly 90% of the word tokens, suggesting that a certain percentage of nouns and verbs appear quite often in the Sinica Corpus. In contrast, the percentage of verbs and nouns in the TMC Corpus is only 45%. Words of the other categories, such as adverbs, pronouns, determinatives, prepositions, and conjunctions, were used significantly more often in conversation than in text.

Table 5. Word Category Distribution.

TMC Corpus	Coverage	Sinica Corpus	Coverage
Verb	23.30%	Noun	53.05%
Noun	22.05%	Verb	36.78%
Adverb	20.01%	Adverb	3.01%
Pronoun	9.98%	Determinative	2.60%
Determinative	6.42%	Foreign words	2.28%
Preposition	5.19%	Adjective	1.24%
Conjunction	4.66%	Conjunction	0.35%
Others	8.39%	Others	0.69%

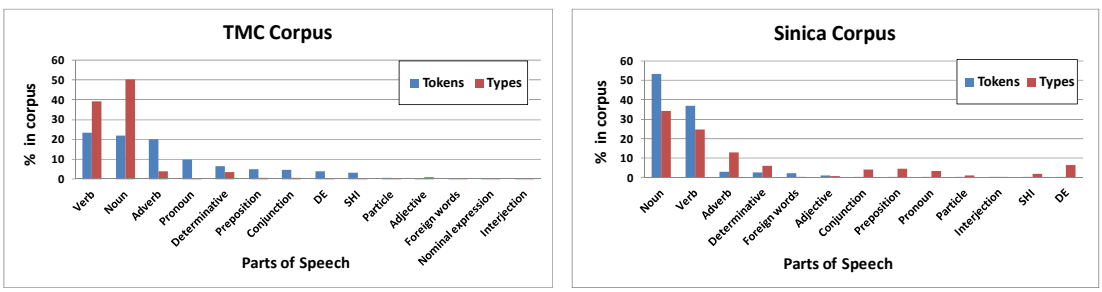

Figure 4. Parts of speech in conversational and text corpus.

The tokens per type of verb and noun in the Sinica Corpus are high because the corpus share of tokens is high and that of types is rather low, as shown in **Figure 4**. This may be due to the topics and the types of the articles included in the corpus, as the Sinica Corpus contains a large number of literary texts. On the contrary, the other word categories cover a much lower share of tokens, but more of types. In the TMC Corpus, a complementary distribution was observed. Verbs and nouns account for wider corpus coverage in terms of types than in terms of tokens. This suggests that different tasks and scenarios of conversations may elicit different

vocabularies. The other word categories, mostly function words, account for more tokens than types. In particular, the use of adverbs is different in conversation and in text. This, to a certain degree, is similar to the distribution found in a comparative study of spoken and written corpora of Swedish (Allwood, 1998). Adverbs, like the other function word categories (conjunctions and prepositions) were used more frequently in the spoken corpus than in the written corpus. Nevertheless, unlike in Taiwan Mandarin, pronouns and verbs were the most frequently produced categories in Swedish text and spoken corpora. The reason may lie in the characteristic of Chinese syntax. Zero anaphora is an often observed phenomenon in Chinese sentences. Therefore, pronouns are often used for addressing people in an interactive communication situation, for instance in conversation. As observed in the comparison of text and conversation, pronouns only make up 0.18% of the overall word tokens in the Sinica Corpus, but 10% in the TMC Corpus.

3.4 Discourse-related Items in the TMC Corpus

Interaction in conversation is often marked by pragmatic indicators, such as prosodic prominence, or by the use of discourse items, such as particles or feedback items. In this regard, conversation clearly differs from text. This section is concerned with corpus coverage of discourse-related items in the TMC Corpus. Compared with ordinary words, discourse items were produced much more frequently. The proportion of the occurrences of ordinary words over those of the discourse items is approximately eight to one in the TMC Corpus. That is, on average, a speech stretch of a length of eight words is accompanied by at least one discourse item. These items mark discourse-relevant positions in conversation, and they usually are produced with distinctive prosodic patterns to indicate the structure of a spoken discourse. With regard to information delivery, they may be considered a kind of redundancy. Their main function is to express the attitudes (particles), the fluency (markers and fillers), and the attention (feedback words) of the speakers. Without these discourse-related items, a conversation would be more like a scripted dialogue.

Table 6. Discourse-related items in conversation.

Groups	Tokens	Types	Tokens per type
Discourse particles	34,842	49	711
Discourse markers	16,516	9	1,835
Fillers/feedback words	6,338	65	98

For academic purposes, we need to investigate these discourse items, because they function as a kind of juncture between concepts and also function as markers of emerging patterns in conversation. As listed in **Table 6**, the tokens per type of discourse markers are 1,835, which is very high compared with ordinary words in the corpus. This suggests that the

performance of automatic speech recognition systems working with conversation can be improved in an economical and efficient way by implementing information and knowledge about the position of these discourse-related items (syntactic or prosodic) and their phonetic representation. Discourse particles are produced more often than the top 1000 word types in the TMC Corpus, 342 tokens per type. The numbers of the distinct types of discourse particles and markers are small, but the tokens per type are high. Furthermore, fillers and feedback words have a limited number of phonetic variants, as their phonetic representations are systematically predictable. Thus, they can be studied in terms of their phonetic forms, pronunciation variations, and their relationship to the contextual information. Feedback words normally mark the structure of speaker turn changes. Automatic detection of the discourse items would significantly enhance the understanding of conversation content and structure.

4. Conclusion

Spoken language is performed differently, given different speaking situations. To understand the lexical capacity of language users, no matter what purposes we have in mind, we need to base our investigations on realistic language data. The ideal corpus of this kind should take into account the versatility of speaker groups, conversation types, and speaking situations. In other words, it needs to be balanced among a variety of sociolinguistic settings. The concept of a balanced corpus for texts needs modification to be used for speech, as a balanced corpus of spoken data should also involve the spontaneous and interactive behavior of the speakers in specific speaking situations. Furthermore, the processing and presentation of speech corpora go beyond the consideration of the meta-data structures of text corpora. The transcribing convention needs to deal with the diversity of spoken phenomena in spontaneous speech. The alignment with the speech signal needs to manually or automatically be conducted to increase the innovative values of speech corpora applications for language technology system and language teaching tools. It is unlikely that the study of lexical coverage based on the Taiwan Mandarin Conversational Corpus represents the capacity of all Taiwan Mandarin speakers in all kinds of speaking situations. Nevertheless, we presented an attempt to provide empirical data for this line of research. With this data, we hope to extend our understanding about the notion how and why humans are capable of conversing by words for communication.

Acknowledgements

The author is grateful to the useful comments provided by two anonymous reviewers of the *International Journal of Computational Linguistics and Chinese Language Processing*. The author also sincerely thanks to the team members who have been working on the corpus data along the years. The study presented in this article is funded by the National Digital Archives Project and the National Science Council under Grant NSC-100-2410-H-001-093.

References

Allwood, J. (1998). Some frequency based differences between spoken and written Swedish. In *Proceedings of the XVIth Scandinavian Conference of Linguistics*, Department of Linguistics, University of Turku.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., & Weiner, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, 24(4), 351-366.

Baayan, R. H. (2001). *Word Frequency Distribution*. Kluwer Academic Publishers. Dordrecht/Boston/London.

Boersma, P. & Weenink, D. (2012). Praat: doing phonetics by computer. http://www.fon.hum.uva/praat 5.3.16.

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.

Chao, Y. R. (1965). *A Spoken Grammar of Chinese*. University of California Press.

Chen, K.-J. & Huang, C.-R. (1996). The SINICA CORPUS": Design methodology for balanced corpora. In *Proceedings of the Eleventh Pacific Asia Conference on Language, Information and Computation*, 167-176.

CKIP. (1998). *The Sinica Corpus 3.0*. The Chinese Knowledge Information Processing Group - technical report 98-04. Academia Sinica. (In Chinese)

Crowdy, S. (1993). Spoken corpus design. *Literary and Linguistic Computing*, 8(4), 259-265.

French, N., Carter, C. W., & Koenig, W. (1930). The words and sounds of telephone conversations. *Bell System Tech Journal*, 9, 290-324.

Gibbon, D., Moore, R., & Winski, R. (Eds.) (1997). *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter.

Howes, D. (1964). Application of the word frequency concept to aphasia. In A. V. S. de Reuck and M. O'Connor, *Disorders of Language* (Ciba Foundation Symposium). London: Churchill, 47-75.

Howes, D. (1966). A word count of spoken English. *Journal of Verbal Learning and Verbal Behavior*, 5(6), 572-606.

Knowles, G. (1990). The use of spoken and written corpora in the teaching of language and linguistics. *Literary and Linguistic Computing*, 5(1), 45-48.

Leech, G., Rayson, P., & Wilson, A. (2001). Word Frequencies in Written and Spoken English - Based on the British National Corpus. Pearson Education Limited.

McCarthy, M. (1999). What constitutes a basic vocabulary for spoken communication? *Studies in English Language and Literature*, 1, 233-249.

Reppen, R. & Ide, N. (2004). The American National Corpus: Overall Goals and the First Release. *Journal of English Linguistics*, 32, 105-113.

Schriffrin, D. (1988). *Discourse Markers*. Cambridge University Press.

Shriberg, E. E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. Doctoral dissertation. Department of Psychology, University of California at Berkeley.

Svartvik, J. & Quirk, R. (1980). *A Corpus of English Conversation*. Lund, Sweden: Gleerup.

Tao, H.-Y. (2009). Core Vocabulary in Spoken Mandarin and the Integration of Corpus-Based Findings into Language Pedagogy. In *Proceedings of the 21st North American Conference on Chinese Linguistics (NACCL-21)*. Vol. 1. Edited by Yun Xiao, 13-27. Smithfield, Rhode Island: Bryant University.

Thorndike, E.L. (1921). The Teacher's Word B*ook*. New York: Teachers College. Columbia University.

Thorndike, E.L. & Lorge, I. (1944). *The Teacher's Word Book of 30,000 Words*. New York: Teachers College, Columbia University.

Tseng, S.-C. (2004). Processing spoken Mandarin corpora. *Traitement automatique des langues*. Special issue: Spoken corpus processing, 45, 89-108.

Xiao, R., Rayson, P., & McEnery, T. (2009). *A frequency dictionary of Mandarin Chinese: Core vocabulary for learners*. Routledge Frequency Dictionaries. London and New York: Taylor and Francis Group.

Wepman, J. M. & Lozar, B. (1973). The most frequently used words of spoken English. *Journal of Psycholinguistic Research*, 2(2), 129-136.

Appendix A: The top 100 words in the core vocabulary

Word	POS	TMC tokens	TMC %	Sinica tokens	Sinica %	Word	POS	TMC tokens	TMC %	Sinica tokens	Sinica %
的	DE	15778	3.89	28582	6.00	上	Ng	1339	0.33	8650	0.18
是	SHI	13999	3.45	84014	1.76	可	D	1337	0.33	8508	0.18
一	Neu	13397	3.30	58388	1.22	爲	VG	1300	0.32	8369	0.18
在	P	7429	1.83	56769	1.19	或	Caa	1296	0.32	8317	0.17
有	V_2	7092	1.75	45823	0.96	好	VH	1273	0.31	8304	0.17
個	Nf	6991	1.72	41077	0.86	等	Cab	1264	0.31	8070	0.17
我	Nh	6705	1.65	40332	0.85	又	D	1197	0.30	8037	0.17
不	D	6677	1.65	39014	0.82	將	D	1161	0.29	7858	0.16
這	Nep	6330	1.56	33659	0.71	後	Ng	1160	0.29	7752	0.16
了	Di	5453	1.34	31873	0.67	因爲	Cbb	1115	0.28	7592	0.16
他	Nh	5301	1.31	30025	0.63	於	P	1030	0.25	7395	0.16
也	D	5260	1.30	29646	0.62	由	P	1001	0.25	7344	0.15
就	D	4827	1.19	29211	0.61	從	P	989	0.24	7303	0.15
人	Na	4694	1.16	24269	0.51	更	D	971	0.24	7298	0.15
都	D	4473	1.10	20403	0.43	被	P	953	0.24	7272	0.15
說	VE	4419	1.09	19625	0.41	才	Da	877	0.22	7266	0.15
而	Cbb	4414	1.09	18452	0.39	已	D	863	0.21	7256	0.15
我們	Nh	4242	1.05	18152	0.38	者	Na	850	0.21	7221	0.15
你	Nh	4100	1.01	17298	0.36	每	Nes	841	0.21	7207	0.15
了	T	3882	0.96	15958	0.33	次	Nf	840	0.21	7087	0.15
要	D	3435	0.85	15955	0.33	把	P	837	0.21	7024	0.15
之	DE	3412	0.84	15893	0.33	三	Neu	834	0.21	6954	0.15
會	D	3398	0.84	14066	0.30	什麼	Nep	832	0.21	6729	0.14
對	P	3173	0.78	13944	0.29	問題	Na	814	0.20	6683	0.14
及	Caa	3124	0.77	13758	0.29	其	Nep	801	0.20	6667	0.14
和	Caa	2932	0.72	13585	0.28	讓	VL	782	0.19	6624	0.14
與	Caa	2832	0.70	13445	0.28	此	Nep	748	0.18	6599	0.14
以	P	2276	0.56	13172	0.28	做	VC	721	0.18	6597	0.14
很	Dfa	2189	0.54	13013	0.27	再	D	716	0.18	6563	0.14
種	Nf	2088	0.52	12263	0.26	所以	Cbb	708	0.17	6529	0.14

中	Ng	2066	0.51	12231	0.26	只	Da	684	0.17	6521	0.14
的	T	1976	0.49	11580	0.24	與	P	665	0.16	6519	0.14
大	VH	1926	0.48	11577	0.24	沒有	VJ	651	0.16	6510	0.14
能	D	1907	0.47	11125	0.23	則	D	646	0.16	6476	0.14
著	Di	1901	0.47	11026	0.23	台灣	Nc	633	0.16	6414	0.13
她	Nh	1869	0.46	10776	0.23	卻	D	630	0.16	6388	0.13
那	Nep	1848	0.46	10740	0.23	地	DE	620	0.15	6329	0.13
上	Ncd	1768	0.44	10619	0.22	並	Cbb	618	0.15	6171	0.13
但	Cbb	1697	0.42	10242	0.21	位	Nf	615	0.15	6015	0.13
年	Nf	1650	0.41	10127	0.21	得	DE	609	0.15	5969	0.13
還	D	1644	0.41	9698	0.20	去	D	604	0.15	5748	0.12
可以	D	1641	0.40	9671	0.20	呢	T	593	0.15	5577	0.12
時	Ng	1633	0.40	9565	0.20	學生	Na	593	0.15	5523	0.12
最	Dfa	1628	0.40	9416	0.20	表示	VE	592	0.15	5504	0.12
自己	Nh	1579	0.39	9069	0.19	到	P	572	0.14	5468	0.11
爲	P	1573	0.39	9026	0.19	公司	Nc	569	0.14	5421	0.11
來	D	1566	0.39	8992	0.19	將	P	568	0.14	5365	0.11
所	D	1518	0.37	8873	0.19	如果	Cbb	563	0.14	5336	0.11
他們	Nh	1500	0.37	8818	0.18	社會	Na	563	0.14	5282	0.11
各	Nes	1454	0.36	8651	0.18	看	VC	562	0.14	5198	0.11