A Taiwan Southern Min spontaneous speech corpus for discourse prosody

Sheng-Fu Wang, Janice Fon

Graduate Institute of Linguistics, National Taiwan University, Taiwan
10617 Le-xue Building, No.1, Sec. 4, Roosevelt Rd., Taipei, Taiwan (R.O.C.)
sftwang0416@gmail.com, jfon@ntu.edu.tw

Abstract

This paper presents a Taiwan Southern Min (Taiwanese) spontaneous speech corpus primarily constructed and annotated for studying discourse prosody. The corpus contains monologue-like speech elicited from interviews. Eight hours of speech contributed by sixteen interviewees, evenly split by gender and age, have been transcribed and annotated. Transcription and the recordings were aligned at the level of syllable with the aid of EasyAlign (Goldman, 2011). Discourse annotation was done by identifying one-verb clausal units and labeling the strength of unit transitions to show the hierarchical structure of discourse using Grosz and Sidner's model (1986). As for prosodic labeling, two major levels of prosodic breaks were identified,along with truncation and prolongation caused by disfluencies and hesitation. The present state of the corpus allows for research on the relationship between acoustic cues, prosodic structure, and discourse organization in unscripted speech.

Index Terms: discourse, spontaneous speech, Taiwanese, Southern Min, prosodic break

1. Introduction

Natural speech is inherently variable. The construction of spontaneous speech corpora is one way of approaching such fascinating variability. This kind of corpora that contain speech collected by a more natural setting provides a greater range of variability than a reading task or other kinds of simple buts designed laboratory setting. Especially in studying the relationship between natural discourse and speech, spontaneous speech is able to reveal phenomena that would otherwise be unavailable to researchers.

The paper presents a spontaneous Taiwan Southern Min speech corpus constructed with the aforementioned objectives in mind. Another important object concerns the target language. Taiwan Southern Min, more commonly known as Taiwanese, is the native language of approximately 70% of the population in Taiwan [1]. Because of the policy that promoted Mandarin Chinese as the official language, Taiwan Southern Min was marginalized in the realm of education, media, and administration. A consequence in linguistic studies was that language resources such as annotated speech corpora in Taiwan Southern Min have been very scarce as compared with resources in Mandarin Chinese. The construction of a spontaneous speech corpus will certainly be an important step to further understandings on relevant theoretical issues on Taiwan Southern Min, as well as providing an important resource for applicational purposes such as speech synthesis and speech recognition.

2. Corpus Construction

The corpus is the continuation of a Mandarin-Min bilingual spontaneous speech corpus started in 2004 [2]. Currently in the annotated dataset, there are eight hours of monologues contributed by sixteen native speakers of Taiwan Southern Min. These speakers can be grouped according to gender and age. For the age of the speakers, the young group contains speakers born in the 1980s and the old group contains speakers born in the 1940s. All of the speakers were from the same region (Taichung, the mid-Taiwan metropolitan area) so that research claims based on the corpus would not be confounded by dialectal differences.

Speech was elicited in the form of an interview in which the interviewer asked the interviewee to talk about his or her personal experiences in childhood, in school, or at work. Marriage, health, and traveling experiences were also common topics. The aim of the interviewer was to elicit longer monologues rather than to engage in a conversation with the interviewee. Whenever the interview became loaded with short turn exchanges between the interviewer and the interviewee, the recording would be not be included in the present dataset.

The currently annotated dataset contains 110873 syllables, 10603 discourse boundaries, and 19433 prosodic breaks. The annotation conventions for discourse and prosodic break will be described in Section 3.

2.1. Transcription

The recordings were transcribed with the Taiwanese-Romanization convention mainly based on the online dictionary constructed by Iu^n [3]. The convention uses both Chinese characters and romanized transcription. Chinese characters are used when it is possible to identify the source characters for a given Southern Min expression. When the characters were not readily identifiable, romanization was used instead.

When the transcription was aligned with the recording in Praat [4], all Chinese characters were romanized. The particularities of the romanization includes using the *h* symbol for plosive aspiration (e.g., *pha* for /p^{h}a/), double *n* for nasalized vowels (e.g., *pinn* for /pĩ/), and *h* for glottal stops at the coda position (e.g., *peh* for /pe?/).

Since Taiwan Southern Min is a tone language, information on lexical tones is provided in the annotation. In addition, since lexical tones in Taiwan Southern Min may be realized as a fixed low tone [5], and the occurrence of such a low tone is not predictable from the transcription, further effort was devoted in identifying syllables with the low tone.

2.2. Syllable alignment

Syllable segmentation and alignment were first automatically conducted with the EasyAlign plug-in [6] and were further manually checked by the first author. Since investigation on the relationship between syllable duration and other linguistic parameters was a primary goal for research based on this corpus, it was crucial to determine the criteria for deciding syllable boundaries, especially at the utterance-final positions. The major criterion taken by the labeler was to put the syllable boundary at the decay of formant structures as shown on Praat spectrograms.

A second trained phonetician labeled 10% of the data and a test of cross-labeler agreement was conducted. Results showed a mean difference of 12.4 ms in terms of the alignment of syllable boundaries, which amounts to 5.7% of mean syllable duration. In addition, 91% of the alignment differences are smaller than 20% of mean syllable duration. This degree of agreement is on par with, or even slightly better as compared with similar tests for the Switchboard Corpus [7], which reports a mean phone-alignment difference of 16.4 ms (19% of mean phone duration), and of the Buckeye corpus [8], which reports that 75% of the phone-alignment differences are smaller than 20% of phone duration.

3. Annotation

3.1. Discourse annotation

Discourse segmentation was done with the transcribed texts of the recordings. The annotation was based on Fon's [9] adaptation of Grosz and Sidner's [10] model on "Discourse Segment Purpose", which identifies the basic intentional units that structured in a hierarchical fashion. In practice, the texts were first segmented into basic discourse units, which are clauses, defined as units that contain one verb, according to the definition of "a simple clause" by a classical study on the functional grammar of Mandarin Chinese [11]. Next, the relationship between clauses was judged according to the level of discourse juncture. Four different levels of "Discourse Boundary Indices (DBI)" were distinguished in this study.

The first level was DBI0, which suggests that the two adjacent clauses describe the same entity or event, thus the boundary between these two is merely a clausal boundary. In the corpus, DBI0 is often used in the following situations: between a matrix and a subordinate clause Examples 1 and 2), two clauses showing parallel syntax (Example 3), between a tag question and its preceding clause (Example 4), between clauses sharing an anaphora (Example 5), and the boundary between two simple discourse units having the relation of cause-effect or topic-comment (Example 6).

(1) [伊講]$_{DBI0}$ ["我嘛欲來"]
"[he said]$_{DBI0}$ ['I also want to come']"

(2) [這我感覺]$_{DBI0}$ [足歡喜ê]
"[it makes me feel]$_{DBI0}$ [very happy]

(3) [可能按呢彼個m̄著]$_{DBI0}$ [啊彼個著按呢]
"[perhaps that was right]$_{DBI0}$ [still that was right]"

(4) [咱也無法度kah伊陪伴伊一世人啊]$_{DBI0}$ [著無?]
"we can't be on her side for her whole life/$_{DBI0}$ right?"

(5) [伊其實本來著無心欲買]$_{DBI0}$ [只是入來趖趖]
"[she didn't really want to buy]/$_{DBI0}$ [(she) just came in and looked around]"

(6) [對以前ê查某來講翁婿著是家己ê天]$_{DBI0}$
[所以一定愛結婚]
"[a husband was like the sky for women in the past]$_{DBI0}$
[so getting married was a must]"

DBI1 refers to the juncture around which the clauses describes about different subtopics or "scenes" within a theme or an episode in narration. In the corpus, DBI1 is often used to label the following cases: an anaphoric change or update (Example 7), a change of aspect or the introduction of a new time reference (Example 8), comments (Example 9), and the boundary between a complex discourse unit to a simple or another complex discourse unit where the units between the boundary have the relation of cause-effect or topic-comment(Example 10).

(7) [因為阮哥哥他們攏讀到高職]$_{DBI0}$
[讀了專科讀了]$_{DBI0}$
[就攏去做工作啊]$_{DBI1}$
[阿本來阮阮阿公的意思是講]...

"[because my brothers they went to the vocational high school]$_{DBI0}$
[after graduating from the vocational high school]$_{DBI0}$
[(they) went to work]$_{DBI1}$
[and originally my grandpa's intention was]..."

(8) [若講交朋友啦]$_{DBI0}$
[我其本上都是專門靠he lah]$_{DBI1}$
[基本上啊我讀二專ê時陣]...

"[when talking about making friends]$_{DBI0}$"
[basically I particularly depended on that]$_{DBI1}$
[basically, when I was studying at the two-year technological college]"

(9) [假那鳥仔放出籠leh]$_{DBI1}$ [足歡喜]
"[just like birds out of their cage]$_{DBI1}$ [very happy]"

(10) [但是我是帶tī農家]$_{DBI1}$
[所以對這leh採棉ê空課嘛lóng真真熟]$_{DBI0}$
[ah實在有落去做]

"[but I lived at the farm]$_{DBI1}$
[so I was quite familiar with the work concerning the silkworms]$_{DBI0}$
[I actually did it]"

DBI2 referred to situations where the boundary clearly differentiates two themes or episodes, yet these themes and episodes are still within a bigger general topic. Example 11 showed one such switch, in which the speaker was still narrating the experiences working in a hospital but changed from the sentiment of the fragility of human-beings to the description of how long he had worked there.

DBI3 was an additional label for handling radical shifts of themes which may be considered boundaries of totally different pieces of monologue or interview within the same recording. This label was often used when the interviewer directed the interviewee to another totally unrelated topic. Example 12 shows a rare case where the jump between topics were initiated by the speaker herself, as she switched from describing her son's working experience to her happy days as a young girl.

(11) [彼陣都一種感覺啦]$_{DBI0}$
[感覺講]$_{DBI0}$
[人原仔是真<MAN 脆弱 MAN>按呢]$_{DBI2}$
[啊彼陣佇遐做做做]$_{DBI1}$
[我會記ê做一兩年吧]

"[at that time there was a feeling]$_{DBI0}$
[(I) felt]$_{DBI0}$
[that human beings are very fragile]$_{DBI2}$
[at the time I worked here (for some time)]$_{DBI1}$
[I remember it was a year or two]"

(12) [有啊]$_{DBI0}$
[逐工轉來啊]$_{DBI1}$
[lóng愛足暗才轉來eN]$_{DBI3}$
[我會感覺哦]$_{DBI0}$
[查某gín仔時代eh足快樂ê]

[yes]$_{DBI0}$
[(he; the speaker's son) came back everyday]$_{DBI1}$
[always came back until it's very late]$_{DBI3}$
[I feel that]$_{DBI0}$
[(I) was very happy when I was a young girl]

The labeling described above is believed to be able to reflect the hierarchical organization of discourse units. The resulting labeling on discourse structure was subsequently annotated and aligned with recordings using Praat [4]. The first author labeled all of the data. A second labeler labeled two of the sixteen transcription files, and the agreement rate was 85%. The discrepancies were discussed and the rest of the labeling were rechecked accordingly by the first author.

3.2. Prosodic annotation

Prosodic units was also annotated in with a ToBI-style system of prosodic breaks. Although there is an prosodic annotation framework of Taiwanese available [12], in their proposal, shown in Table 1, the level below the intonation phrase (IP) is the Tone Sandhi Group (TSG), whose occurrence is defined by rule-governed tonal alternations. This kind of definition is very different from what is commonly used for defining or describing a level of prosodic unit, such as the perception of a certain pitch movement or acoustic cues such as final lengthening and final pitch-lowering[13, 14, 15]. Also, the occurrence of TSG boundaries can almost regularly be predicted from syntax [16, 17], with a very low degree of freedom if the syntactic structure is to be held unaltered. It makes the inclusion of TSG as a level of prosodic phrasing doubtful.

Table 1: TW-ToBI break indices [12]

b4	intonation phrase boundary, either utterance-finally or -medially
b3	tone sandhi group (TSG) boundary
b3m	percept of TSG boundary without sandhi tone
b2m	base tone without percept of the TSG ending
b2	ordinary "word-internal" syllable boundary
b1	resyllabification
b0m	syllable fusion

Thus, a new labeling scheme was devised, as shown in Table 2. This new scheme focuses on two things: The first focus is the differentiation of the intonation phrase and a lower level of prosodic unit. This differentiation makes sure that discourse boundaries corresponding to the same level of prosodic boundaries are investigated. The second focus is the annotation of hesitation, truncation, and abrupt stops.

Table 2: A proposed Taiwan Southern Min BI tagset

4	intonation phrase boundary, either utterance -finally or -medially
4-	utterance-final boundary without obvious IP boundary cues
3	a lower level of boundary, utterance-medially
3p	hesitation that gives the percept of a major prosodic boundary
2p	hesitation that does not give the percept of a major prosodic boundary
1p	truncation and abrupt stop

Figure 1 and Figure 2 show examples of a level 4 and level 3 break in the proposed break index framework. The difference between these two types of breaks can be clearly seen from Figure 2, where the syllable at level 3 break has lowered F0 not lengthened as a level 4 break does. Figure 3 shows a prolonged syllable perceived as hesitation, labeled with 3p. Figure 4 labeled two cases of 1p: the former showed an abrupt stop followed by an immediate repair, and the latter showed a segmental deletion.

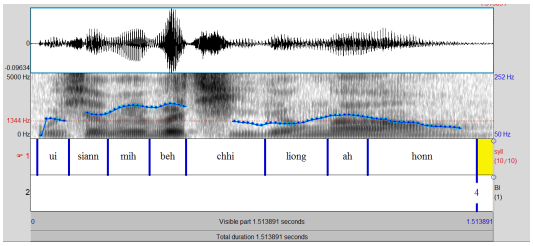

Figure 1: A BI4 (intonational phrase boundary) example; "the reason why I raise fish"

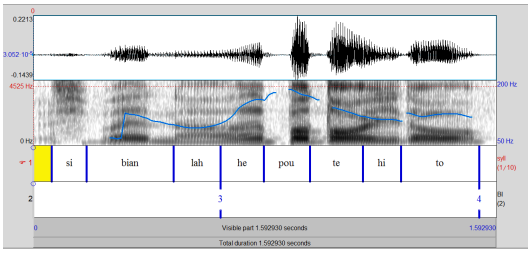

Figure 2: A BI3 boundry (tentative label for a prosodic break smaller than an intonational phrase boundary but bigger than word boundary) followed by a BI4 boundary; "No, the glove puppetry"

A second labeler annotated the prosodic breaks of 10% of the corpus and the agreement with the author was calculated with the kappa statistic [18]. The kappa statistic on boundary placement was 0.86. When both labelers agreed on placing a BI, the kappa statistic for BI category agreement was 0.6. These values are comparable with the report on the interlabeler agreement on ToBI labeling on the Switchboard corpus [19], which yielded a kappa statistic of 0.75 for pitch accent placement, 0.67

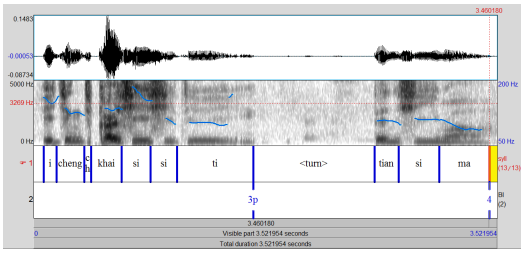

Figure 3: A BI3p boundary that labeled prolongation perceived as hesitation; "in the past it was on..... TV"

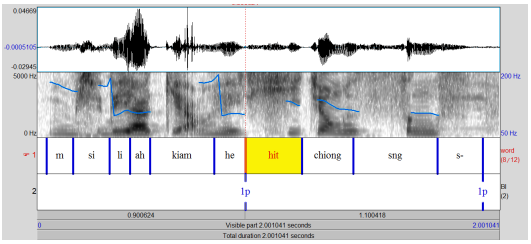

Figure 4: Two 1p boundaries that labeled abrupt stop and segmental deletion: "Not the salted plum; that kind of..."

for phrasal accent placement, 0.58 for boundary tone placement, and 0.51 for pitch accent choice, and 0.48 for phrasal accent choice. The only statistic that exceeds our current agreement rates was the kappa value for boundary tone choice (0.71). [20] also examined interlabeler agreement on Break Index in the ToBI framework, which yielded a kappa value of 0.75 for phrasal boundary placement and 0.67 for phrasal boundary size, slightly exceeding the value obtained in the current study. Table 3 presents the agreement matrix of break indices.

Table 3: Agreement matrix of break indices (Column headings indicate labels assigned by the author and row headings are labels assigned by the second labeler)

	1p	2p	3	3p	4	NO	Sum
1p	189	10	6	10	78	64	357
2p	0	18	10	10	5	23	66
3	5	10	61	1	49	81	207
3p	1	30	5	153	55	13	257
4	43	5	44	25	868	103	1088
NO	17	8	91	2	34	8226	8378
Sum	255	81	217	201	1089	8510	10353

4. Conclusions & Prospects

The present state of the corpus allows for research on the relationship between acoustic cues, prosodic structure, and syntactic/discourse structure. In addition to further annotation on larger amounts of data, more dimensions of annotation such as POS tagging and the identification of Tone Sandhi Group boundaries will also be implemented to make the corpus a more valuable resource for a wider varieties of research topics on Taiwan Southern Min.

5. References

[1] Government Information Office, *The Republic of China Yearbook 2012*. Kwang Hwa Pub. Co., 2012.

[2] J. Fon, "A preliminary construction of taiwan southern min spontaneous speech corpus," Tech. Rep. NSC-92-2411-H-003-050, National Science Council, Taiwan, 2004.

[3] U.-G. Iu^n, "Taiwen huawen xianshangzidian jian-zi jishu ji shiyongcingxing tantao [Construction and utilization of Taiwanese-Chinese online dictionary]," in *Proceedings of the 3rd International Conference on Internet Chinese Education*, pp. 132–141, 2003.

[4] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program], version 5.1. 44," 2010.

[5] U.-J. Ang, *Taiwan Helaoyu shengdiao yanjiu (Research on Taiwanese Tones)*. Taipei: Zili wanbao, 1985.

[6] M.-H. Chen, J.-P. Goldman, H.-h. Pan, and J. Fon, "Easyalign: an automatic phonetic alignment tool under praat," in *Proceedings of the Workshop on New Tools and Methods for Very-Large-Scale Phonetics Research*, vol. 2011, pp. 109–112, 2011.

[7] S. Greenberg, J. Hollenback, and D. Ellis, "Insights into spoken language gleaned from phonetic transcription of the switchboard corpus," in *International Conference on Spoken Language Processing*, pp. S32–35, Citeseer, 1996.

[8] W. Raymond, M. Pitt, K. Johnson, E. Hume, M. Makashay, R. Dautricourt, and C. Hilts, "An analysis of transcription consistency in spontaneous speech from the buckeye corpus," in *Proceedings of the International Conference on Spoken Language Processing*, pp. 1125–1128, 2002.

[9] Y.-J. J. Fon, *A cross-linguistic study on syntactic and discourse boundary cues in spontaneous speech*. PhD thesis, The Ohio State University, 2002.

[10] B. Grosz and C. Sidner, "Attention, intentions, and the structure of discourse," *Computational linguistics*, vol. 12, no. 3, pp. 175–204, 1986.

[11] C. Li and S. Thompson, *Mandarin Chinese: A functional reference grammar*. Univ of California Pr, 1981.

[12] S. Peng and M. Beckman, "Annotation conventions and corpus design in the investigation of spontaneous speech prosody in taiwanese," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[13] M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, "The original ToBI system and the evolution of the ToBI framework," *Prosodic typology: The phonology of intonation and phrasing*, pp. 9–54, 2005.

[14] S. Godjevac, "Transcribing serbo-croatian intonation," *Prosodic typology: The phonology of intonation and phrasing*, pp. 146–171, 2005.

[15] J. J. Venditti, "The J_ToBI model of Japanese intonation," *Prosodic typology: The phonology of intonation and phrasing*, pp. 172–200, 2005.

[16] R. L. Cheng, "Tone sandhi in taiwanese," *Linguistics*, vol. 6, no. 41, pp. 19–42, 1968.

[17] M. Chen, "The syntax of xiamen tone sandhi," *Phonology Yearbook*, vol. 4, pp. 109–149, 1987.

[18] J. Carletta, "Assessing agreement on classification tasks: the kappa statistic," *Computational linguistics*, vol. 22, no. 2, pp. 249–254, 1996.

[19] T. Yoon, S. Chavarria, J. Cole, and M. Hasegawa-Johnson, "Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI," in *Proceedings of the International Conference on Spoken Language Processing*, pp. 2729–2732, Nara Japan, 2004.

[20] M. Breen, L. C. DiLLey, and J. KraeMer, "Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch)," *Corpus Linguistics and Linguistic Theory*, vol. 8, no. 2, pp. 277–312, 2012.