Investigating Discourse Segmentation in Taiwan Southern Min Spontaneous Speech

Laurent Prévot[1,2]
[1]CNRS & MEAE, CEFC
[2]CNRS & Aix Marseille Université, LPL
Taipei, Taiwan
laurent.prevot@cefc.com.hk

Sheng-Fu Wang
Institute of Linguistics
Academia Sinica
Taipei, Taiwan
sftwang@gate.sinica.edu.tw

Abstract

In recent years, discourse segmentation has received increased attention; however the majority of studies have focused on written genres and languages with abundant linguistic resources. This paper investigates discourse segmentation of a spontaneous speech corpus in Taiwan Southern Min. We compare fine-tuning a Language Model (LLM) using two approaches: supervised, taking advantage of a high-quality annotated dataset, and weakly-supervised, which requires only a small amount of manual labeling. The corpus used here is transcribed in both Chinese characters and romanized script. This allows us to assess the impact of the written form on the discourse segmentation task. Moreover, the dataset includes manual prosodic break labeling, allowing an exploration of the role prosody can play in contemporary discourse segmentation systems grounded in LLMs. In our study, the supervised approach outperforms weak-supervision; the character-based version demonstrates better scores compared to the romanized version; and prosodic information proves to be an interesting source to increase discourse segmentation performance.

1 Introduction

Discourse segmentation consists in breaking down texts or conversations into functional units that better corresponds to participants' intentions than sentences or simple speech activity chunks. We will use the term discourse unit (DU) (Asher and Lascarides, 2003) to designate a minimal speech act or communicative unit. Each DU corresponds roughly to a clause-level content that denotes a single fact or event.

While the segmentation of discourse units (DUs) in written documents has received a lot of attention from the discourse and NLP community, the same cannot be said for the segmentation of spontaneous speech. In this study, we approach the segmentation of discourse units in a corpus of spontaneous speech in Taiwan Southern Min.

Southern Min is a sino-tibetan language spoken by over 50 million people, and includes Taiwan Southern Min, which is one of the official language of Taiwan. We take advantage here of an existing discourse segmented corpus of spoken interviews for running discourse segmentation experiments.

We develop DU segmenters based on different principles and evaluate their performance. More precisely, we compare fine-tuning an LLM with hand labeled data vs. employing a data programming approach (Ratner et al., 2017) that requires only a fraction of annotated data. While fine-tuning LLMs for language well represented in the LLM training data proved to be a very efficient solution (Gravellier et al., 2021; Prevot et al., 2023), it remains to be seen whether this approach is relevant for languages, particularly their spontaneous speech variants, less represented in the training data. Finally, we investigate the impact of using either romanization or Chinese characters in our dataset, as well as the potential contribution of prosody.

2 Related Work

In recent years, there has been a renewed interest in discourse parsing and discourse unit segmentation within the NLP community. As in other subdomains, Large Language Models have proven highly beneficial and allowed to reach unprecedented scores for these tasks. However, discourse segmentation within these deep learning approaches has been applied to only a few langauges, until the recent initiative of DISRPT campaigns started (Zeldes et al., 2019, 2021; Braud et al., 2023). The work conduced within the framework of these campaigns has equipped the community with a set of powerful tools and frameworks to perform DU segmentation using these contemporary approaches.

Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024), pages 50–63
March 21, 2024 ©2024 Association for Computational Linguistics

As discussed in Braud et al. (2023), even for written genres, discourse segmentation performance drops in languages other than English and when gold sentences are not given, due to sentence segmenters being far from perfect (Braud et al., 2017). Considering spontaneous conversational speech, the related tasks of dialogue-act segmentation and tagging yiels various interpretation regarding the definition of base units. For instance, some models explain that dialogue acts being multi-functional, several segmentations can be considered depending on the aspects of dialogue being considered at the time of segmentation (Petukhova et al., 2011).

A recent trend involves approaching discourse segmentation with sequential models over contextual embeddings (Wang et al., 2018; Muller et al., 2019). Turning specifically to spontaneous speech discourse segmentation, (Gravellier et al., 2021) applied a weak-supervision approach (Ratner et al., 2017) and reached an f-score of 73.7 while having access to gold turn segmentation. More specifically, manual heuristic rules, including some rules exploiting the discourse segmentation model trained on a written dataset (Muller et al., 2019), were created to annotate noisily the entire dataset. This noisy data was then used to fine-tune an LLM, BERT (Devlin et al., 2018) in that case. In Prevot et al. (2023), a larger amount of manual annotation allowed to compare fine-tuning with larger amount of training data and a weakly-supervised approach. For this French dataset, it was concluded that more than 7000 annotated DUs were required in the supervised training approach to beat the weakly-supervised approach (f-score: 70.6). When more data was used, supervised fine-tuning reached slightly higher scores (f-score: 73.9). These f-score results are $10 - 15\%$ than the scores obtained on written genress, which is expected as sentence splitters leveraging punctuation provide substantial assistance for discourse unit segmentation. In speech, particularly spontaneous interactional speech, pauses are useful but are by far less reliable in predicting discourse units since they are involved in many other dimensions and are subject to significant inter-individual variability. Recently Metheniti et al. (2023)[1], an improvement over Muller et al. (2019) has been developed, allowing to reach new state-of-the-art results for discourse segmentation in various languages. Our paper reuses the technical framework of this paper.

Segmenting speech into Discourse and Prosodic units has been the focus of numerous studies across various languages, including high-resource languages like English (Hirschberg and Grosz, 1992; Hirschberg and Nakatani, 1996), Dutch (Swerts, 1997), French or Mandarin (Degand and Simon, 2009; Prévot et al., 2015) as well as low-resource languages (Mettouchi and Vanhove, 2021). Discourse-prosodic interface research has also been developed for better understanding turn-taking mechanisms (Hu and Degand, 2023; Botinis et al., 2007). The deep connection between discourse and prosody has led researchers to explore prosodic cues for discourse tasks with some success (Pierrehumbert and Hirschberg, 1990; Shriberg et al., 2000). However, to our knowledge, there are no studies in which modern LLM-based systems described above, which achieve high scores based solely on transcripts, have benefited from incorporating acoustic-prosodic cues. An interesting attempt was made in (Gravellier et al., 2021), which validated the weak-supervision approach exploiting silent pauses among other elements, but the results did not improve with the inclusion of other acoustic-prosodic features. This is likely due to (i) the already high scores obtained from text alone, which would require cues coming from other sources to yield very high precision; and (ii) to the challgenge of automatic extracting reliable prosodic cues, such as speech rate, pitch or even intensity, from conversational speech.

Discourse Studies on Southern Min (and related language like Hakka or Cantonese) have focused on final particles (Lien, 1988; Li, 1999; Fung, 2000; Chappell, 2019), which can carry an interesting range or semantic and pragmatic functions. Moreover, there have been specific corpus studies examining discourse markers in Taiwan Southern Min (Chang, 2002, 2008; Chang and Hsieh, 2017). However, to the best of our knowledge, there has been no attempt to automatically segment discourse units in this language.

Additionally, there have been specific corpus studies examining (Chang, 2002, 2008; Chang and Hsieh, 2017). However, to the best of our knowledge, there has been no attempt to automatically segment discourse units in this language.

[1] Code at https://github.com/phimit/jiant/

3 Dataset

3.1 Base data

The discourse segmentation data used in this paper comes from an 8-hour corpus of monologue-like spontaneous speech elicited in sociolinguistic interviews as part of a larger project that collected Min-Mandarin bilingual speech recordings all over Taiwan between 2004 and 2010 (Wang and Fon, 2013; Fon, 2004). This subset of the corpus, also used in phonetic studies on phenomena including pre-boundary lengthening (Wang, 2023, 2022; Wang and Fon, 2012) and tone sandhi (Chen, 2018), contained speech materials from 16 speakers, who each contributed around 30 minutes of recording. The speakers were evenly split in gender and two age groups (old and young). At the time of recording, the old speakers were between 50-65 years old, and the young speakers were between 20-35 years old. Due to the original recording setup, the transcripts only focused on speech from the interviewee, with the interviewer's turns being labeled with a 'turn' token. The transcripts follow the convention used in a dictionary [2] administered by the Ministry of Education in Taiwan, along with a romanized version. The transcripts were aligned with the recordings at the syllable level using EasyAlign (Goldman, 2011) with manual corrections from a trained phonetician. During the manual correction process, pauses annotation was incorporated in the transcripts that are used in this study. In addition to pauses, the corpus also contains annotations on prosodic breaks, with a main goal of identifying the presence of two levels of breaks (intonational phrases and intermediate phrases), as well as breaks resulted in from hesitations and disfluencies. Data from two of the speakers were used to calculated cross-labeller agreement (kappa: 0.86). We observe that although done completely independently discourse and prosodic units exhibit a relationship : 45% of the prosodic breaks are also discourse breaks while 82% of the discourse breaks also correspond to a prosodic break.

Due to the lack of widely available text-processing tools in this language, dictionary-based method was used to perform word segmentation (maximal length matching) and POS tagging, the latter of which follows a multihot format, i.e., a word that is ambiguous between multiple POS tags according to the dictionary is annotated as '1' for all those tags.

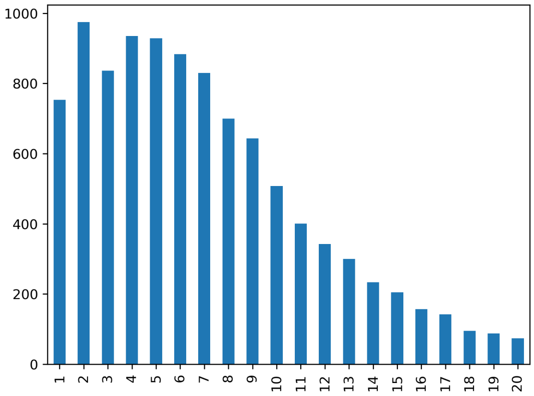


Figure 1: DU lengths in tokens.

The corpus contains 88.5K words at the word level with pause (#) and specific interviewer turn symbols included.

3.2 Discourse Segmentation Annotation

The corpus contains annotation of discourse units, which are defined as units that contain a verb and its core arguments, a criterion that is also used in other studies on the interaction between discourse and prosody (e.g., (Chen and Tseng, 2019; Prévot et al., 2015)). Crucially, discourse annotation in this corpus was performed independently from the recordings, i.e., the annotators only saw the transcripts, with turn information but no precise timing information, when they performed the task. Similarly to the prosodic labeling, two annotators labeled transcripts from two of the speakers for examination of interlabeller agreement (kappa: 0.96), and one annotator labeled the remaining transcripts. See Table 1 for examples of discourse units.

Disfluencies were not segmented apart and were instead included within discourse units. Discourse labellers had access to gold turn segmentation but were not told to use them systematically. As a result a few discourse units manually labeled span over more than one turn.

Taking a more quantitative perspective, the distribution of the annotated discourse units lengths in terms of tokens is provided in 1. We can see a fairly balanced distribution of lengths that are shorter than 10 tokens with a mean of 7.5 tokens per discourse unit. Truly conversational corpora

[2] https://sutian.moe.edu.tw/zh-hant/

char:	[其實 # 我 相信]	[別人 會使] #	[咱 就 一定 會使]
roman:	[ki5-sit8 # goa2 siong-sin3]	[pat-lang5 e7-sai2]	[lan2 to it-teng7 e7-sai2]
gloss:	[actually (pause) I believe]	[others can]	[we PART must can]
trans:	[‘actually I believe’]	[‘(if) others can (do it)’]	[‘we must be able to (do it as well)’]

Table 1: Examples of three discourse units. Note how the pause (#) may occur within a discourse unit

tend to present a different bimodal distribution with a mode of very short units (made of 1 token) corresponding to feedback and back-channels and a second mode of units made of 4-6 tokens. The dataset here is a corpus of interviews for which only the interviewee is transcribed. While being truly spontaneous, this explains why there are less extremely short interactional units as well why the mode of the distribution includes longer lengths than purely dialogic genres.

4 Methodology / Experiments

The corpus includes interviews of 16 speakers. We made 8 folds composed of two speakers each and ran a cross-validation over the 8 folds with different test / dev / train splits. Given our corpus, this is a method that maximizes the distance between training and testing data.

Two main approaches are evaluated for segmenting automatically our dataset : (i) directly fine-tuning a LLMs with all the data at our disposal (in a supervised way) (*Supervised* setting), (ii) create a noisily annotated datasets thanks to manual heuristic rules (See Figure 2) and a model to combine them.

More specifically, we used ROBERTA (Liu et al., 2019) and the framework fine-tuning it was DISCUT (Metheniti et al., 2023), grounded in JIANT environment (Pruksachatkun et al., 2020).

The weak-supervision framework uses SKWEAK (Lison et al., 2021) rather than SNORKEL (Ratner et al., 2017). SKWEAK natively allows the model to exploit the sequential nature of our task. On the technical side, SKWEAK relies on SPACY (Honnibal and Montani, 2017) documents. In order to keep all the relevant information (timing, pos-tags, prosody labels) linked to the tokens and to use them in the labeling rules, we made use of SPACY extensions attributes.

In the weak supervised approach, we use SKWEAK's ability to build a generative model from noisy labels provided by the labeling rules. SKWEAK allows to choose an HMM to perform this sequence labeling task. While this approach can be adopted without annotated data, a small development set is useful for testing and crafting the heuristic labeling rules. We can decide more efficiently which manual rules should be retained, dropped or improved thanks to the metrics that are computed on the development set. Besides precision, recall and f-score, *overlaps* and *conflicts* (with other rules) metrics are also useful to take decisions over the usage of these rules (See table 2).

name	label	conflict	precision	recall
#_begpos	BDU	0.14	0.86	0.19
turn	BDU	0.16	0.84	0.11
beg_char	BDU	0.25	0.75	0.21
conj	BDU	0.36	0.64	0.24

Table 2: Profiles for a few labeling rules

To summarize, the weakly supervised approach is performed as follows:

1. write the labeling rules (See Figure 2) ;
2. apply and evaluate them on the `dev set` (iterate with the previous step until satisfied with labeling rules profiles on `dev set`) (See profiles in Table 2);
3. apply the labeling rules to the `train set`;
4. fit the HMM SKWEAK (rules aggregation) model;
5. apply the resulting model to the `test set`.

For the time being, the labeling rules crafted are extremely simple. They are using (i) pause duration and turn information; (ii) frequent tokens present at discourse boundaries; (iii) POS-tags overrepresented at discourse boundaries. Moreover, manually annotated prosodic units boundaries are included in the dataset and we use them for some experiments. As mentioned above, POS-tags are encoded in a multihot format. The labeling rules

```
def pause_and_begin_char(doc):
    for idx, token in enumerate(doc):
        if idx > 0:
            if (doc[idx-1].text == '#') and (doc[idx-1]._.dur > PAUSE)
                    and (doc[idx].text in BEGIN_CHAR):
                yield idx,idx+1,'BDU'
            else:
                yield idx,idx+1,'ABS'
        else:
            yield idx,idx+1,'BDU'
```

Figure 2: Labelling Function example (pause combined with a DU-initiating character)

exploiting POS are formulated accordingly to this ambiguous situation.

Characters vs. letters The corpus we are working with includes two versions of the transcription: characters and romanization (as seen in example 1). All our experiments were realized in both written forms.

Prosodic boundaries This corpus comes with prosodic break expert manual annotations. For the gold dataset, we created two versions of the dataset : one without any kind of prosodic information; and one with a special token corresponding to the presence / absence of a prosodic break. This special token was added to the transcript in all datasets (train / test / dev).

5 Results

The results comparing the general approach are presented in figure 3; the one related to the impact of the written form used are in figures 4 and 5 and the results of the prosody experiments are visualized in 6. All the numbers can be checked in Annex 3.

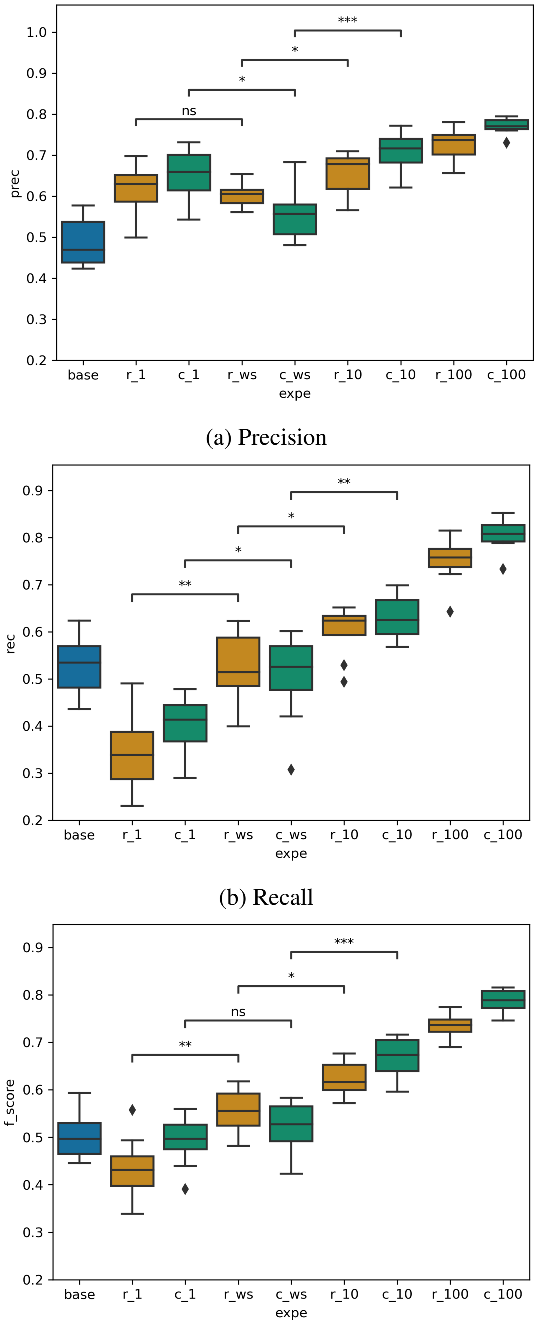
Supervision or weak-supervision Our results[3] (presented in Figure 3) shows that our weak-supervision approach remains behind from the supervised approach. This is true with large amount of manually annotated training data (∼70K tokens)[4] but the difference is already significant with smaller amounts of training data (∼7K tokens) for precision, recall and f-score (P:70.8/R:63.0/F:66.7). Weak supervision does better only if extremely limited amount of training data is available (∼700 tokens).

Which base units? The results of the experiments show that different written forms (characters vs. romanized) for the corpus yielded signicantly different results. The difference between the two versions of the corpus lies in the fact some romanized tokens correspond to several characters (e.g., ‘ah’ corresponds to ‘啊’, an utterance-initial/final particle, and ‘矣’, a sentence-final particle and perfective aspect marker; ‘e5’ corresponds to ‘的’, a possessive marker and sentence-final particle, ‘个’, a classifier, and ‘鞋’, a noun for ‘shoe’.), while there are also some, but much less, characters that correspond to different romanizations (e.g., ‘嘛’ correspond to ‘ma7’, which means ‘also’, and ‘mah’, a final particle). This situation conduced us to propose several hypotheses. First of all, when there is not a lot of fine-tuning data, having less symbol types can help to get faster a robust model. When more annotated data is available, having more specific symbols should bring better results by revolving some ambiguities. However, a second fact to consider is that the LLM we are fine-tuning (ROBERTA) includes Mandarin Chinese but not Southern Min. We therefore hypothesized that the character version should have an advantage when very little amount is provided since the base symbols are present in the model to fine-tune while the romanized symbols featuring tone digits should be something completely new for the model.

The results presented in Figure 4 show an advantage to character based corpus with large amount of fine-tuning data (Characters: 77.0/80.5/78.7 ;

[3] In all the paper, the significance labels included in the figures are corresponding to *p-values* of a *t-test* done on the folds of the experiment. A difference between two conditions is said to be significant (*/**/***) if t-testing the two series of values coming from the folds for both conditions, yielded the corresponding threshold p-values (0.05 / 0.01 / 0.001).

[4] For characters, supervised approach gives an f-score of 78.7 (p:77.0/r:80.5) while weak supervision only reaches a 52.0 f-score (p:55.7/r:50.4).

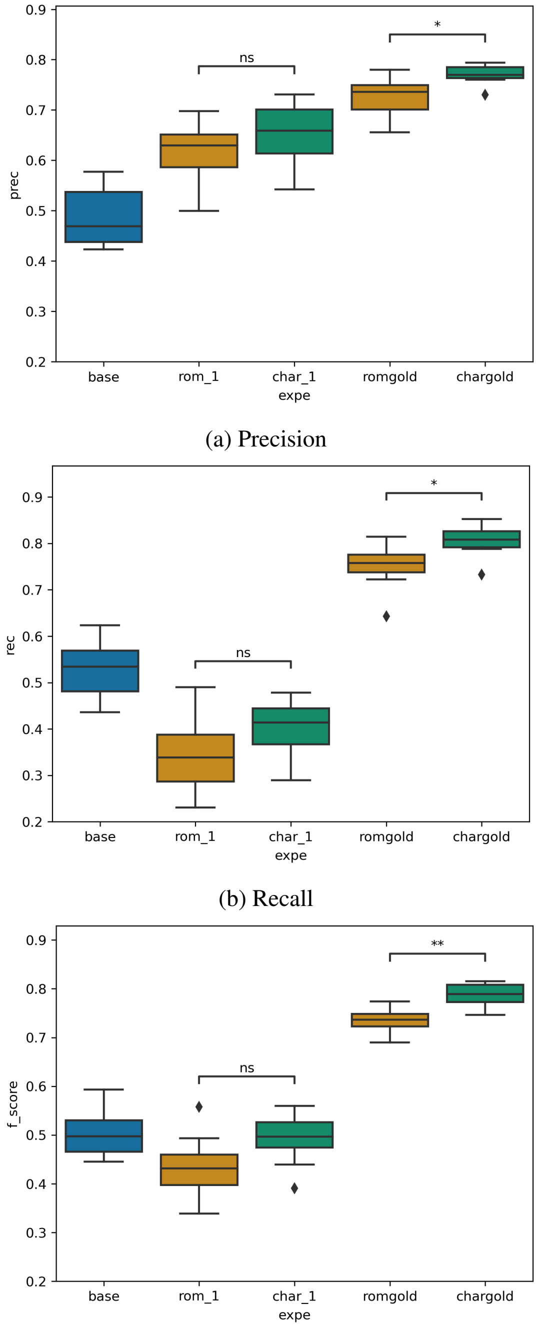


(a) Precision

(b) Recall

(c) F-score

Figure 3: Supervised vs. Weakly-supervised. *blue : 200ms pause baseline; orange : romanized; green: characters. From left to right _1:1% training data (∼700 toks), _10:∼7K toks), _100:∼70K toks)*

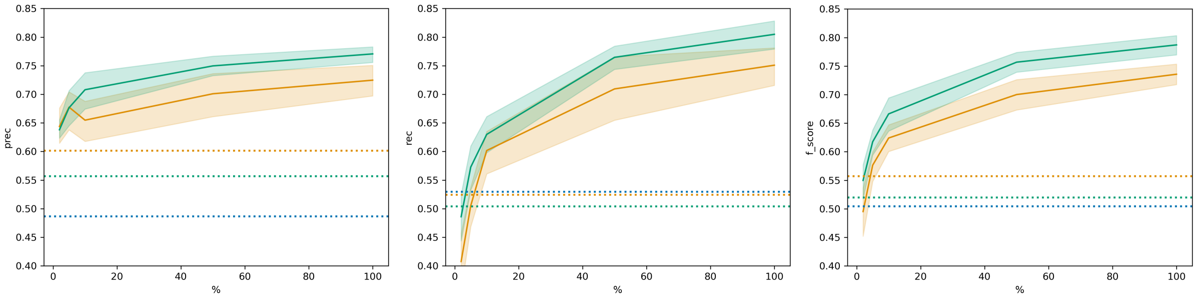


(a) Precision

(b) Recall

(c) F-score

Figure 4: Characters vs. Romanized. *blue: 200ms pause baseline; orange: romanized; green: characters. From left to right _1:1% training data (∼700 toks), _100:∼70K toks)*



(a) Precision (b) Recall (c) F-score

Figure 5: Amount of training data. *orange: romanized corpus ; green: character version. From 1% training data (∼700 toks) to 100% (∼70K toks). Dotted lines, blue: baseline, green and orange : weak supervision*

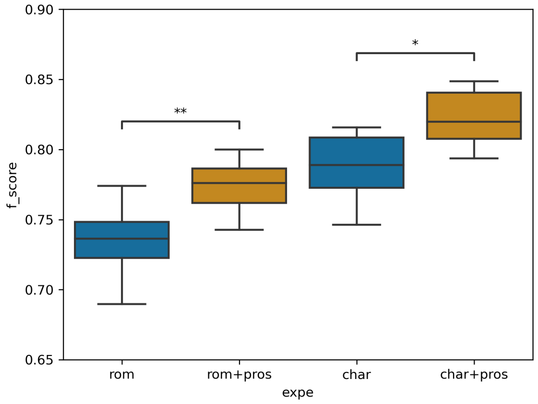


Figure 6: Adding prosody. F-score

Romanized: 72.5/75.1/73.6). It seems to be also the case when little amount of data is provided but this difference did not reach statistical significance. There also seems to be some complexities where we could expect to find a sweet spot for the romanized version (a little data for fine-tuning but not a lot, see the precision and recall with 5% and 10% of training data on figure 5) but the numbers do not allow to conclude on this result.

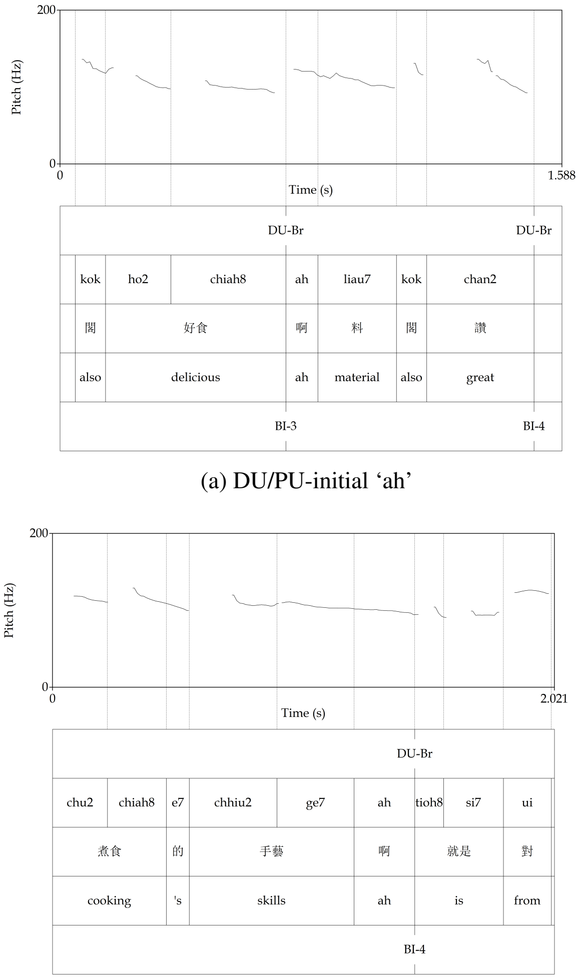
Potential help from prosody Prosody information used in this study had been manually added. As explained above, this prosodic annotation is however completely independent from the discourse segmentation. From a linguistic perspective, prosody should help in segmenting discourse units in speech since segmentation is one of the linguistic function of prosody (Swerts, 1997; Hirschberg and Grosz, 1992; Degand and Simon, 2009; Di Cristo, 2013). However, the recent work of (Gravellier et al., 2021), realized in a similar framework as ours, did not show the benefit of adding prosodic-acoustic cues for performing discourse segmentation. This was based however on automatic acoustic extraction. Given the data available to us, we decided to test whether "gold" prosodic segmentation would help on discourse segmentation performance. More precisely, every token in our dataset carries the information of whether it is at the beginning of a prosodic unit or not.

The base model we used did not allow for an enrichment at the token level. We therefore translated the prosodic information into a token. More precisely, for each start of labeled prosodic unit we inserted a rare character in the transcript. The figure 6 illustrates the statistically significant benefit of adding prosodic information for the characters and romanized versions of the corpus. The increase for the character version was +4.5,+2.5 and +3.5 for precision, recall and f-score respectively. These increases might seem modest but one should remember that pause duration and turn information was already taken into account before exploiting these prosodic labels.

6 Error Analysis

To further understand how our models could be improved we performed a detailed qualitative error analysis of the various models output.

(1) is an example where the model trained on gold and WS show the same segmentation error: While the gold annotation does not segment this sequence into two DUs, the models put a boundary after the sentence-final particle ‘oh’ and a pause. It is a representative example on the overuse of pause as a segmentation cue, especially for the WS-trained model. It also shows that the human annotator has a stronger tendency to only segment DUs with a main verb (thus ‘reversely my only friend oh’ is not a DU) while also neglecting potential disfluencies and false starts (‘reversely is’). It



(a) DU/PU-initial 'ah'

(b) DU/PU-final 'ah'

Figure 7: Illustration of prosodic help to discourse unit segmentation: (a) The particle 'ah' being used as a DU-initial marker is coincided with an intermediate phrase break (BI-3) signaled by pitch reset, i.e., higher f0 at 'ah'. (b) The particle is DU-final and exhibit lengthening and continued f0 declination with the preceding syllable, both of which are characteristics of an intonational phrase boundary (BI-4).

is worth noting that while the literal word sequence contains 'reversely is', the whole phrase has the same interpretation as 'reversely'. The presence of complex adverbs and/or discourse markers is likely another reason that this task is challenging for the models.

(1) 'On the other hand, my boyfriend oh he would still gone to see me' (GEN: genitive marker; PART: a marker similar to ba5 in Mandarin ba construction.)

  a. Gold annotation: [ah reversely is # reversely is I GEN boy friend oh # he still would go PART me see]

  b. Gold & WS-trained: [ah reversely is # reversely is I GEN boy friend oh #] [he still would go PART me see]

(2) is another example where the gold-trained model oversegmented a DU that was viewed by the human annotator as a noun and a relative clause ('The boyfriends that I had').

(2) 'The boyfriends that I had I always didn't marry them'

  a. Gold annotation (and WS-trained): [I self have GEN boy friend all all marry no success]

  b. Gold trained: [I self have GEN boy friend] [all all marry no success]

Finally, (3) shows an example of how gold-trained and WS-trained segmentation may differ from the gold annotation in distinct ways. The gold annotation has a DU boundary between the main clause and the tag question, the former containing some disfluencies. The model trained on gold annotation did not recognize the boundary with the tag question and instead put a boundary before the word 'like this' (an2-ne), which reflects the fact that an2-ne is a discourse marker that can occur in clause-initial and clause-final positions. The model trained on WS data, on the other hand, did not put a DU boundary for the entire sequence (thus having an error of under-segmentation before 'you know not'), as there was no pause nor words that have a strong tendency to start a DU in the corpus.

(3) 'At that time, walking still didn't require tiptoeing, you know?' (hyphen-connected units denote a word in TSM).

a. Gold annotation: [Then walking still does-not like this does-not require tiptoeing] [you know not]

b. Gold-trained: [Then walking still does-not] [like-this does-not require tiptoeing you know not]

c. WS-trained: [Then walking still does-not like-this does-not require tiptoeing you know not]

7 Discussion and Future Work

In this paper, we applied state-of-art techniques of discourse segmentation to a dataset of Taiwan Southern Min. We compared supervised and weakly supervised approaches. Moreover the linguistic information included in the original dataset allowed us to test some hypotheses along the way. We tested whether (i) it was easier to segment with the character-based or romanized version of the corpus ; and (ii) prosodic gold labels could help these new models of discourse segmentation.

An important overall result is that the approach employed (fine-tuning a sequence-to-sequence model) performs extremely well on this Taiwan Southern Min corpus, a language not included in the base Language Model (LLM) used. This is an important result with regard to the applicability of such approaches to low-resource languages for this task. The longer term goal of this work is to apply the best model we can build to a much larger corpus of Taiwanese interviews. The results obtained enable us to try to replicate existing studies on discourse-prosody interface in spontaneous speech, which have relied solely on manually annotated data.

Getting into the comparison of the two approaches tested, we should remind here that the scores obtained with gold annotations should be taken as a top line for the weak supervision approach. Indeed, the amount of manual gold segmentation for this corpus is substantial and does not aligh with the typical scenario for adopting a weak-supervision approach. With this consideration in mind, we observe that the weakly supervised approach failed to produce comparable results to the supervised setting. This can be attributed on the one hand to the supervised approach yielding highly competitive results through fine-tuning with only about 10% of our full amount of annotated data (corresponding $7K$ tokens, 700 discourse units); and on the other hand to the relatively low performance of our weakly supervised model. However, this does not negate the potential interest of weak supervision. Our current rules are rudimentary, primarily using simple pauses, tokens information and ambiguous POS-tags. We intend to enhance these labeling rules in several directions: (i) using a real POS-tagger that would reduce ambiguity ; (ii) developing more sophisticated labeling rules to address phenomena specific to spontaneous speech, such as disfluencies.

Regarding the comparison between the character-based and romanized versions of the corpus, the clear conclusion is that the character version consistently yields better results regardless of the amount of fine-tuning data provided. This could be attributed to both the benefit of lower ambiguities of characters over romanized version and to the presence of Mandarin data in RoBERTa.

Regarding prosody, this study has shown that, in line with linguistic predictions and previous computational models, but contrary to recent findings on this task, prosodic information can indeed help in discourse unit segmentation. The next obvious step is to automatize the extraction of relevant acoustic features that approximate efficiently the manual annotations we had in this stydy. From the primary prosodic features identified in (Shriberg et al., 2000) for English, excluding the ones already exploited by our pause and turn related rules, we identify (i) pitch differences across the discourse unit boundary, and (ii) duration of phones and rhymes preceding the decision point.

Acknowledgments

The authors would like to thank Philippe Muller and Pierre Magistry for side discussions related to the content of the paper. This study was supported by the Institute of Linguistics, Academia Sinica (LING-113-018-01), Taiwan's National Science and Technology Council (NSTC-112-2410-H-001-098-MY2) as well as by the Institut Convergence ILCB (ANR-16-CONV-0002) and the ANR-Funded project SUMM-RE (ANR-20-CE23-0017).

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Antonis Botinis, Aikaterini Bakakou-Orphanou, and Charalabos Themistocleous. 2007. Mutlifactor analysis of discourse turn in greek. In *16th International Congress of Phonetic Sciences*, pages 1341–44.

Chloé Braud, Ophélie Lacroix, and Anders Søgaard. 2017. Does syntax help discourse segmentation? not so much. In *Conference on Empirical Methods in Natural Language Processing*, pages 2432–2442.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

Miao-Hsia Chang. 2002. Discourse functions of anne in taiwanese southern min. *Concentric: Studies in English Literature and Linguistics*, 28(2):85–115.

Miao-Hsia Chang. 2008. Discourse and grammaticalization of contrastive markers in taiwanese southern min: A corpus-based study. *Journal of pragmatics*, 40(12):2114–2149.

Miao-Hsia Chang and Shu-Kai Hsieh. 2017. A corpus-based study of the recurrent lexical bundle ka li kong 'let (me) tell you'in taiwanese southern min conversations. *Chinese Language and Discourse*, 8(2):174–211.

Hilary Chappell. 2019. Southern min. *The mainland Southeast Asia linguistic area*, pages 176–233.

Alvin Cheng-Hsien Chen and Shu-Chuan Tseng. 2019. Prosodic encoding in mandarin spontaneous speech: Evidence for clause-based advanced planning in language production. *Journal of Phonetics*, 76:100912.

Mao-Hsu Chen. 2018. *Tone Sandhi Phenomena in Taiwan Southern Min*. University of Pennsylvania.

Liesbeth Degand and Anne Catherine Simon. 2009. On identifying basic discourse units in speech: theoretical and empirical issues. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (4).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Albert Di Cristo. 2013. *La prosodie de la parole*. De Boeck Superieur.

J Fon. 2004. A Preliminary construction of Taiwan Southern Min spontaneous speech corpus. Technical Report NSC-92-2411-H-003-050.

Roxana Suk-Yee Fung. 2000. *Final particles in standard Cantonese: semantic extension and pragmatic inference*. The Ohio State University.

J.P. Goldman. 2011. EasyAlign: an automatic phonetic alignment tool under Praat. In *Proceedings of Interspeech 2011: 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31*, pages 3233–3236.

Lila Gravellier, Julie Hunter, Philippe Muller, Thomas Pellegrini, and Isabelle Ferrané. 2021. Weakly supervised discourse segmentation for multiparty oral conversations. In *Proceedings of EMNLP 2021*.

Julia Hirschberg and Barbara Grosz. 1992. Intonational features of local and global discourse structure. In *Proceedings of the DARPA workshop on Spoken Language Systems*. Association for Computational Linguistics.

Julia Hirschberg and Christine H Nakatani. 1996. A prosodic analysis of discourse segments in direction-giving monologues. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 286–293.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

Junfei Hu and Liesbeth Degand. 2023. The conversational discourse unit: Identification and its role in conversational turn-taking management. *Dialogue & Discourse*, 14(2):83–112.

Ing Cherry Li. 1999. *Utterance-final particles in Taiwanese: A discourse-pragmatic analysis*. Crane Publishing Company.

Chinfa Lien. 1988. Taiwanese sentence-final particles. In Robert L. Cheng and Shuanfan Huang, editors, *The structure of Taiwanese: A modern synthesis*, pages 209–240. The Crane Publishing Taipei.

Pierre Lison, Jeremy Barnes, and Aliaksandr Hubin. 2021. skweak: Weak supervision made easy for nlp. *arXiv preprint arXiv:2104.09683*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Eleni Metheniti, Chloé Braud, Philippe Muller, and Laura Rivière. 2023. DisCut and DiscReT: MELODI at DISRPT 2023. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 29–42, Toronto, Canada. The Association for Computational Linguistics.

Amina Mettouchi and Martine Vanhove. 2021. Prosodic segmentation and cross-linguistic comparison in corpafroas and cortypo: Corpus-driven and corpus-based approaches.

Philippe Muller, Chloé Braud, and Mathieu Morey. 2019. Tony: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 115–124. Association for Computational Linguistics.

Volha Petukhova, Laurent Prévot, and Harry Bunt. 2011. Multi-level discourse relations between dialogue units. In *Proceedings 6th joint ACL-ISO workshop on interoperable semantic annotation (ISA-6), Oxford*, pages 18–27.

Janet Pierrehumbert and Julia Bell Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In *Intentions in communication*. MIT press.

Laurent Prevot, Julie Hunter, and Philippe Muller. 2023. Comparing methods for segmenting elementary discourse units in a French conversational corpus. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 436–446, Tórshavn, Faroe Islands. University of Tartu Library.

Laurent Prévot, Shu-Chuan Tseng, Klim Peshkov, and Alvin Chen. 2015. Processing units in conversation: A comparative study of French and Mandarin data. *Language and Linguistics*, 16(1):69–92.

Yada Pruksachatkun, Phil Yeres, Haokun Liu, Jason Phang, Phu Mon Htut, Alex Wang, Ian Tenney, and Samuel R. Bowman. 2020. jiant: A software toolkit for research on general-purpose text understanding models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 109–117, Online. Association for Computational Linguistics.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access.

Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech communication*, 32(1-2):127–154.

Marc Swerts. 1997. Prosodic features at discourse boundaries of different strength. *The Journal of the Acoustical Society of America*, 101(1):514–521.

Sheng-Fu Wang. 2022. The interaction between predictability and pre-boundary lengthening on syllable duration in taiwan southern min. *Phonetica*, 79(4):315–352.

Sheng-Fu Wang. 2023. Boundary Strength and Predictability Effects on Durational Cues at Tone Sandhi Group Boundaries in Taiwan Southern Min. In *Proceedings of the 20th International Congress of Phonetic Sciences*.

Sheng-Fu Wang and Janice Fon. 2012. Durational cues at discourse boundaries in taiwan southern min. In *Speech Prosody 2012*.

Sheng-Fu Wang and Janice Fon. 2013. A taiwan southern min spontaneous speech corpus for discourse prosody. *The Proceedings of Tools and Resources for the Analysis of Speech Prosody, Aix-en-Provence, France*, pages 20–23.

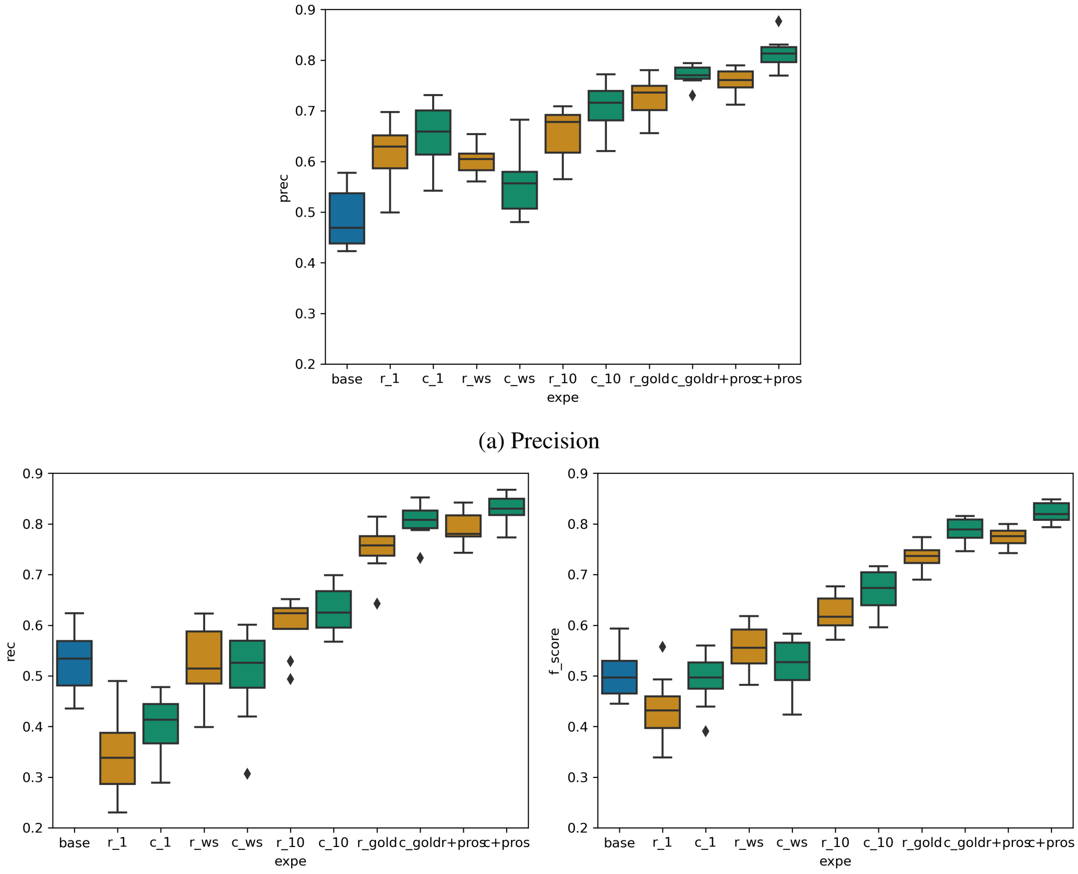
Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 962–967, Brussels, Belgium. Association for Computational Linguistics.

Amir Zeldes, Debopam Das, Erick Galani Maziero, Juliano Antonio, and Mikel Iruskieta. 2019. The disrpt 2019 shared task on elementary discourse unit segmentation and connective detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 97–104.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene, editors. 2021. *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*. Association for Computational Linguistics, Punta Cana, Dominican Republic.

A Appendix

A.1 Global Results



(a) Precision

(b) Recall

(c) F-score

Figure 8: Global Results *blue: baseline, orange: romanized corpus ; green: character version*

	prec mean	prec std	rec mean	rec std	fscore mean	fscore std
pause baseline (200ms)	0.486618	0.060169	0.529578	0.068400	0.504385	0.050271
super. rom (700 toks)	0.616545	0.061804	0.344490	0.081643	0.435640	0.067387
super. char (700 toks)	0.652257	0.063328	0.398917	0.065834	0.490842	0.053958
weakly super. rom	0.601497	0.031159	0.524181	0.077477	0.557128	0.047371
weakly super. char	0.556877	0.064321	0.503797	0.098992	0.519769	0.055981
super. rom (7K)	0.654762	0.054972	0.601636	0.058013	0.624031	0.036572
super. char (7K)	0.707989	0.049716	0.629861	0.049157	0.666265	0.046354
super. rom (70K)	0.724710	0.040760	0.750888	0.052945	0.735763	0.028225
super. char (70K)	0.770644	0.020731	0.804883	0.036518	0.787142	0.025453
super. rom (70K) + pros	0.757477	0.027094	0.792695	0.034534	0.774099	0.020699
super. char (70K) + pros	0.814579	0.031807	0.829729	0.029347	0.821556	0.020996

Table 3: Global Results

A.2 Tokens and POS lists used in the labelling rules

A.2.1 POS list

```
BEGIN_POS = ['interjection']
END_POS = ['interjection', 'onomatopoeia', 'particle']
NON_BEGIN_POS = ['interrogative', 'locative', 'numeral', 'onomatopoeia', 'quantifier']
NON_END_POS = ['adposition', 'conjunction', 'numeral', 'pronoun']
```

A.2.2 Romanized token lists

```
BEGIN_UNI_ROM = ['tan7-si7', 'li5-chhiann2', 'sou2-i2', 'henn', 'ran2m-houm']
END_UNI_ROM = ['lah', 'bo', 'mah', 'neh', 'nia5', 'm']
BEGIN_BI_ROM = ['ah chit-ma2',  'ah na7', 'henn ah', 'li2 e7', 'ah i', 'in-ui7 li2',
               'sou2-i2 gun2', 'ah ma7', 'sou2-i2 goa2', 'ah cho3', 'tan7-si7 goa2',
                'ah si7','ah m7-koh','henn goa2','oh he','ah hit-chun7','ah chiah',
                'tioh8 bo']
END_BI_ROM = ['bo5 lah', 'ni5 ah', 'u7 ah', 'e5 lah', 'ho2 chiah8', 'bo5 ah','ah lah',
              'tioh8 ah', 'si5-chun7 honn', 'lah honn', 'henn ah', 'an2-ne lah',
            'goa2 kam2-kak', 'khi3 ah', 'kam2-kak kong2', 'an2-ne nia5', 'e5 an2-ne',
             'koe3 ah', 'tioh8 lah', 'ho2 ah', 'e5 oh', 'chai-iann2 kong2', 'e5 neh',
              'kang5-khoan2 ah', 'ho2 lah', 'an2-ne honn', 'tioh8 bo']
```

B Labelling Rules

B.1 More examples

```
def very_long_pause(doc):
    for idx, token in enumerate(doc):
        if idx > 0:
            if doc[idx-1].text in PAUSE_TOK and doc[idx-1]._.dur > VERY_LONG_PAUSE:
                yield idx,idx+1,'BDU'
            else:
                yield idx,idx+1,'ABS'
        else:
            yield idx,idx+1,'BDU' #beginning of doc
```

```
def begin_pos(doc):
    for idx, token in enumerate(doc):
        if idx > 0:
            for cat in string_to_list(doc[idx]._.pos_list):
                if cat in BEGIN_POS:
                    yield idx,idx+1,'BDU'
            yield idx,idx+1,'ABS'
        else:
            yield idx,idx+1,'ABS'
```

B.2 Labeling Functions profiles (Romanized)

	annotator	label	conflict	precision	recall	f1
1	non_end_pos	NO	0.028	0.991	0.252	0.401
2	non_begin_pos	NO	0.112	0.970	0.070	0.130
3	cluster_rom_neg	NO	1.000	0.700	0.001	0.002
5	pause_ending_bi_rom	BDU	0.109	0.927	0.048	0.092
6	pause_begin_pos	BDU	0.112	0.888	0.082	0.151
7	begin_bi_rom	BDU	0.121	0.888	0.090	0.163
8	pause_begin_bi_rom	BDU	0.121	0.879	0.048	0.091
9	pause_endrom	BDU	0.200	0.875	0.033	0.064
10	turn	BDU	0.158	0.842	0.111	0.196
11	beginrom	BDU	0.180	0.839	0.172	0.286
12	extreme_pause	BDU	0.181	0.826	0.116	0.204
13	pause_beginrom	BDU	0.181	0.819	0.064	0.119
14	cluster_rom_pos	BDU	0.200	0.800	0.008	0.015
15	endrom	BDU	0.318	0.773	0.016	0.032
16	very_long_pause	BDU	0.263	0.741	0.144	0.241
17	long_pause	BDU	0.417	0.588	0.235	0.335
18	pause_end_pos	BDU	0.463	0.551	0.148	0.233
19	ending_bi_rom	BDU	0.490	0.530	0.101	0.170
20	conjunction	BDU	0.494	0.525	0.128	0.205
21	pause	BDU	0.490	0.514	0.336	0.406
22	short_pause	BDU	0.583	0.424	0.520	0.467
23	begin_pos	BDU	0.597	0.410	0.160	0.230

Table 4: Label Functions profiles for Romanized version