International Journal of

Computational Linguistics & Chinese Language Processing

中文計算語言學期刊

A Publication of the Association for Computational Linguistics and Chinese Language Processing

This journal is included in THCI, Linguistics Abstracts, and ACL Anthology.

易繫辭曰上古結繩而治後世聖人易之以書契百官以治萬民以察說文敘曰蓋文字者經藝之本宣教明化之始前人所以垂後後人所以識古故曰本立而道生知天下之至賾而不可亂也教化既萌文心雕龍則謂人之立言因字而生句積句而成章積章而成篇篇之彪炳

Vol.26 No.1 June 2021 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

International Journal of

Computational Linguistics & Chinese Language Processing

Aims and Scope

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) is an international journal published by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). This journal was founded in August 1996 and is published four issues per year since 2005. This journal covers all aspects related to computational linguistics and speech/text processing of all natural languages. Possible topics for manuscript submitted to the journal include, but are not limited to:

- Computational Linguistics
- Natural Language Processing
- Machine Translation
- Language Generation
- Language Learning
- Speech Analysis/Synthesis
- Speech Recognition/Understanding
- Spoken Dialog Systems
- Information Retrieval and Extraction
- Web Information Extraction/Mining
- Corpus Linguistics
- Multilingual/Cross-lingual Language Processing

Membership & Subscriptions

If you are interested in joining ACLCLP, please see appendix for further information.

Copyright

Cover

Calligraphy by Professor Ching-Chun Hsieh, founding president of ACLCLP
Text excerpted and compiled from ancient Chinese classics, dating back to 700 B.C.
This calligraphy honors the interaction and influence between text and language

International Journal of
Computational Linguistics and Chinese Language Processing
vol. 26, no. 1, June 2021

Contents

Computational Linguistics and Chinese Language Processing
Vol. 26, No. 1, June 2021, pp. 1-16

The NTNU Taiwanese ASR System for Formosa Speech Recognition Challenge 2020

Fu-An Chao[*], Tien-Hong Lo[*], Shi-Yan Weng[*],

Shih-Hsuan Chiu[*], Yao-Ting Sung[+#], and Berlin Chen[*]

Abstract

This paper describes the NTNU ASR system participating in the Formosa Speech Recognition Challenge 2020 (FSR-2020) supported by the Formosa Speech in the Wild project (FSW). FSR-2020 aims at fostering the development of Taiwanese speech recognition. Apart from the issues on tonal and dialectical variations of the Taiwanese language, speech artificially contaminated with different types of real-world noise also has to be dealt with in the final test stage; all of these make FSR-2020 much more challenging than before. To work around the under-resourced issue, the main technical aspects of our ASR system include various deep learning techniques, such as transfer learning, semi-supervised learning, front-end speech enhancement and model ensemble, as well as data cleansing and data augmentation conducted on the training data. With the best configuration, our system obtains 13.1 % syllable error rate (SER) on the final-test set, achieving the first place among all participating systems on Track 3.

Keywords: Formosa Speech Recognition Challenge, Deep Learning, Transfer Learning, Semi-supervised Training

1. Introduction

Due to the rapid developments of deep learning, deep neural network (DNN) based techniques have enjoyed widespread adoption in the automatic speech recognition (ASR) community.

[*] Department of Computer Science and Information Engineering,
National Taiwan Normal University, Taipei, Taiwan
E-mail: {fuann, tienhonglo, 60947007S, shchiu, sungtc, berlin}@ntnu.edu.tw
[+] Department of Educational Psychology and Counseling,
National Taiwan Normal University, Taipei, Taiwan
[#] Institute for Research Excellence in Learning Sciences,
National Taiwan Normal University, Taipei, Taiwan

ASR has also found its applications in many different areas, ranging from interactive voice response (IVR) services and personal assistants, for which people can interact with the machine naturally by using their own voices, meeting transcription, speech translation to speech summarization. Nowadays, some top-of-the-line ASR systems can even reach the performance level of professional human annotators in English, a dominant language in the world. However, in real-world scenarios, there exist some languages that are resource-poor or even endangered. For example, although both Mandarin (a.k.a. Pǔtōnghuà and Huáyǔ) and Taiwanese (a.k.a. Taiwanese Hokkien, Hoklo, Taigi, Southern Min and Min-Nan) are Chinese dialects spoken by large populations of people, the latter is underexplored and has far less ASR training data made publicly available than the former, which causes the performance of Taiwanese ASR systems to fall short of expectations. Furthermore, distinct from Mandarin, Taiwanese has a wide variety of pronunciation traits that can be attributed to the influences from disparate languages like Formosan, Dutch, Japanese, and among others ("Taiwanese Hokkien," 2021). Despite there are still 70% of the population in Taiwan who use Taiwanese to communicate, most of the people, young generations in particular, have only limited vocabulary and cannot speak Taiwanese fluently. Therefore, in addition to the continued promotion of the Taiwanese language and the preservation of its associated culture in the 21st century, how to empower Taiwanese ASR applications in daily life, like voice command control and automatic TV show subtitling and IVR, to name just a few, remains to be of prime importance.

This paper describes the NTNU ASR system participating in the Formosa Speech Recognition challenge 2020 (FSR-2020) supported by the Formosa Speech in the Wild[1] project (FSW). Figure 1 outlines the major components of our system submitted to FSR-2020. The high variety existing in the pronunciation characteristics of Taiwanese and the varying noise-contaminated test conditions of the FSR-2020 datasets make this challenge intrinsically much more difficult. In the setting of the Track 3 competition, the output hypotheses of an ASR system have to be tonal syllable sequences, with a tone index, ranging from 1 to 9, attached to each syllable. Apart from the training data provided by the organizer, all participants were allowed to build their systems with additional data outside the FSR-2020 training dataset. Instead of recourse to extra in-domain speech data, we stick to conducting Taiwanese ASR on a resource-scarce assumption. To this end, we explore the joint use of several ASR modeling strategies, including data augmentation, transfer learning, semi-supervised training and model ensemble. In addition, to alleviate the negative effects of ambient noise and reverberation that may mix with the test utterances of Track 3, speech enhancement is also applied to generate augmented data for training the acoustic models.

[1] https://sites.google.com/speech.ntut.edu.tw/fsw

Finally, our system with the best configuration takes the first place among all participating systems on Track 3.

The remainder of this paper is organized as follows: Section 2 sheds light on our main contributions and the strategies that were employed, from front-end processing to back-end acoustic and language modeling. Section 3 presents the details of the experimental setup, results and discussion. Finally, we conclude the paper and envisage future research directions in Section 4.

2. Strategies for Building a Taiwanese ASR System

In this section, we present our main strategies for building a Taiwanese ASR system for FSR-2020, which consist of two preprocessing procedures, viz. lexicon augmentation and data cleansing, training data augmentation, front-end speech enhancement, acoustic modeling and language modeling. Each of the aforementioned components will be elaborated on in the following subsections, respectively.

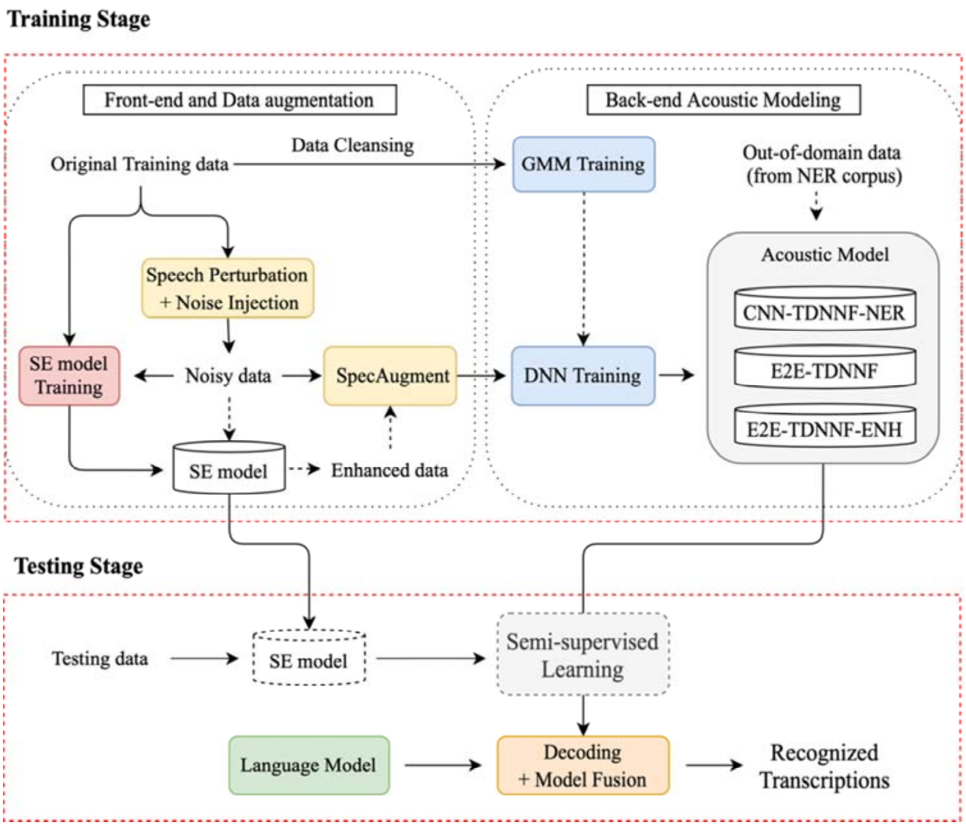

Figure 1. An overview of the NTNU system for FSR-2020.

2.1 Lexicon Augmentation and Data Cleansing

First of all, through a careful inspection of the training dataset and baseline recipe[2] provided by the FSR-2020 organizer, we noticed there were a few syllable patterns in the given lexicon missing their pronunciations. In order to make the lexicon more complete, we used a greedy approach to performing lexicon augmentation. Specifically, we first enumerated all distinct tonal syllables that appeared in the reference transcripts of the training dataset and then augmented the lexicon with those tonal syllables that were not found in the lexicon. As a side note, we excluded none-Taiwanese lexical patterns that appeared in the transcripts of the training dataset, such as English words and proper nouns (e.g., Google), from consideration in this study.

The acoustic model of our ASR system was built with a hybrid deep neural network and hidden Markov model (DNN-HMM) structure, which employed a DNN in place of Gaussian mixture model (GMM) for modeling the state emission probabilities in a traditional GMM-HMM structure. Hybrid DNN-HMM acoustic models have shown to be significantly superior than the conventional GMM-HMM acoustic models on many ASR tasks. It is arguable that Hybrid DNN-HMM acoustic models still have to resort to GMM-HMM acoustic models to obtain good forced-alignment information for better estimation of their corresponding neural network parameters. Inspired by this practice, the GMM-HMM acoustic model of our best system was trained with the audio segments that were screened out from speech training dataset with high recognition confidence scores generated by an existing hybrid DNN-HMM system. As we shall see later, the empirical ASR results confirm this intuitive data-cleansing therapy.

2.2 Data Augmentation

The training dataset for the Track 3 of FSR-2020 was provided by the organizer, which consisted of about 50 hours of Taiwanese utterances. This amount of dataset would be insufficient when training a hybrid DNN-HMM acoustic model for Taiwanese ASR. To enrich the speech training data and increase the robustness of our ASR system for the Track 3 competition, we thus set out to leverage various data augmentation strategies based on different label-preserving transformations.

In addition to utterance-level speed perturbation (Ko *et al.,* 2015) used in our baseline system, we also adopted other data augmentation methods, including spectrogram augmentation and noise injection. As we shall see in the experimental section, these data augmentation strategies collectively lead to promising results that further push the limits of our ASR system.

---

[2] https://github.com/t108368084/Taiwanese-Speech-Recognition-Recipe

2.2.1 Spectrogram Augmentation

Apart from the augmentation strategies that operate in the waveform domain (Ko *et al.,* 2015), feature-based augmentation that is conducted in the feature-space domain (e.g., spectrum or spectrogram) is another active line of research for acoustic modeling in past few years. One of the most celebrated feature-space augmentation methods adopted for acoustic modeling is vocal tract length perturbation (VTLP) (Jaitly & Hinton, 2013). VTLP employs a linear warping transformation along the frequency bins, simulating the effect of altering the vocal tract lengths of speakers that produce the training utterances. More recently, SpecAugment (Park *et al.,* 2019) has drawn much attention from the ASR community. With the inspiration from computer vision (CV), SpecAugment treats the spectrogram of an utterance as an image, which first performs warping along the time axis (time-warping) and then masks blocks of consecutive time and frequency bins in different axes (time-frequency masking). The whole operations of SpecAugment jointly lead to considerable word error rate reductions on several benchmark ASR tasks.

In this paper, we made use of the component "spec-augment-layer" of the Kaldi toolkit (Povey *et al.,* 2011) along with speed perturbation to generating label-preserving, augmented data for training our acoustic model. Note here that "spec-augment-layer" consists of only time and frequency masking operations. This is probably because the time-warping operation is conceptually similar to speed perturbation conducted in the waveform domain, but costs a larger amount of computation and may not lead to substantial improvements (Park *et al.,* 2019).

2.2.2 Noise Injection

To alleviate the deteriorating effects of time- or frequency-varying noise when testing our system in an unseen environment, we also injected different types of noise into the training speech utterances, which were compiled from a few online-available noise datasets (Snyder *et al.,* 2015) (Thiemann *et al.,* 2013) (Dean *et al.,* 2010) (Saki *et al.,* 2016) (Saki & Kehtarnavaz, 2016). In doing so, we can not only increase the diversity of training data but also prevent the hybrid DNN-HMM acoustic model from encountering the overfitting problem. Specifically, we randomly selected a signal-to-noise-ratio (SNR) ranging from -5 to 15 dB when contaminating each speed-perturbed utterance with a certain type of noise, totally creating a 6-fold augmentation of the training dataset for estimating the acoustic model. In addition, these corrupted utterances were also used to train our front-end speech enhancement component (see Section 2.3 that follows).

2.3 Front-end Speech Enhancement

In a wide variety of realistic situations, the input to an ASR system might be noise-contaminated speech. As a solution to mitigate the undesirable noise-contamination effects, speech enhancement (SE) is arguably a crucial modeling paradigm to improve noise robustness of acoustic modeling. Particularly, time-domain SE methods have drawn much attention from both the academic and commercial sectors in past few years and have exhibited outstanding noise-reduction performance on many ASR tasks.

As such, we used the fully-convolutional time-domain audio separation network (Conv-TasNet) (Luo & Mesgarani, 2019) as a preprocessing component for noise suppression, which was originally proposed for the speech separation task intended to separate an input mixture to individual speech signals. Conv-TasNet has shown superior performance over many frequency-domain approaches. The main architecture of Conv-TasNet is composed of an encoder, decoder and masking network, jointly processing a raw waveform signal in an end-to-end manner. As we shall see later, with the adoption of Conv-TasNet, we can obtain marked improvements in both front-end and back-end ASR evaluations.

2.4 Acoustic Modeling

The DNN component of our hybrid DNN-HMM acoustic model involves several layers of factorized time-delay neural network (TDNNF) (Povey *et al.*, 2018), optionally prepended by several layers of convolutional neural network (CNN). Such a pairing of neural networks is denoted by CNN-TDNNF hereafter. TDNNF is viewed as an effective extension to TDNN (time-delay neural network), with the purpose of obtaining better modeling performance and meanwhile reducing the number of parameters by factorizing the weight matrix of each TDNN layer into a product of two low-rank matrices. As a side note, it is worth mentioning that our hybrid DNN-HMM acoustic model can also be estimated with the so-called flat-start, end-to-end training setup suggested by (Hadian *et al.*, 2018) (denoted by E2E-TDNNF hereafter). This setup facilitates the training of a hybrid DNN-HMM acoustic model without resort to any previously trained acoustic models and forced-alignment information, while the estimated model can still work pretty well especially in noisy test scenarios.

In addition, the objective function for training the acoustic model is lattice-free maximum mutual information (LF-MMI) (Povey *et al.*, 2016)

$$\mathcal{F}_{\text{LFMMI}} = \sum_{i=1}^{N} \log \frac{P(\mathbf{O}_i|L_i)^k P(L_i)}{\sum_L P(\mathbf{O}_i|L)^k P(L)} \tag{1}$$

where $\mathbf{O}_i$ and L_i are the acoustic feature vector sequence and the corresponding phone sequence of the i-th training utterance, k is a weighting factor, and $P(L_i)$ is the phone N-gram language model probability.

2.5 Language Modeling

For language modeling, we first adopted the SRILM toolkit (Stolcke, 2002) to train *N*-gram based language models, for which both the *Good-Turing* (Gale, 1995) and *Kneser-Ney* (Ney & Essen, 1991) *N*-gram smoothing methods were considered. In our experiments, we observed that using a four-gram language model yielded considerable improvements than a tri-gram language model in terms of both the perplexity and ASR error reductions. Also worth mentioning is that both these language models were trained solely on the text corpus provided by the organizer.

3. Experiments

3.1 Experimental Setup

Table 1. The statistics of TAT-Vol1 and NER corpus.

TAT-Vol1 (Taiwanese)			
	Speakers	Utterances	Duration
Train	80	22,605	41 hours
Pilot-test	-	2,617	5 hours
Final-test	-	5,663	10 hours
NER (Taiwanese Mandarin)			
	Speakers	Utterances	Duration
Vol1-2-3	-	57,387	360 hours

In the FSR-2020 challenge, the training dataset released by the organizer for developing our ASR systems is TAT-Vol1 (Liao *et al.,* 2020), which is a publicly-available Taiwanese speech corpus. We made use of the whole TAT-Vol1 corpus as our training dataset (about 41 hours) and evaluated our system and various modeling approaches with the pilot-test dataset (about 5 hours; released by the organizer for the warm-up evaluation). To further utilize off-the-shelf audio data, we adopted NER corpus (Liao *et al.,* 2017) into this work, which contains about 360 hours Taiwanese Mandarin speech. The detailed statistics of the corpus are summarized in Table 1, and all the ASR systems with different modeling approaches were developed with the Kaldi toolkit (Povey *et al.,* 2011).

In order to verify the effectiveness of our modeling approaches, we evaluated our systems with two distinct metrics for front-end SE and back-end ASR, respectively. In the front-end experiments, we will evaluate our SE component with the scale-invariant signal-to-noise ratio (SI-SNR) metric, which has been previously shown to be closely related to recognition error reduction (Kinoshita *et al.,* 2020): the higher the SI-SNR score the better the ASR performance. On the other hand, we will use syllable error rate (SER) followed by the rules suggested by (Liao *et al.*, 2020) to evaluate our back-end ASR systems with different modeling approaches. All the SER results depicted in the following experiments will take into account the correctness of the tone index attached to each syllable.

3.2 Experiments on Data Cleansing and Lexicon Augmentation

Table 2. SER (%) results on the pilot-test dataset with different baseline acoustic models.

Acoustic Model	Parameters	Data Cleansing	Lexicon Augmentation	SER (%)
TDNN-F	17M	-	-	19.21
TDNN-F	17M	×	-	17.16
TDNN-F	17M	×	×	17.15
TDNN-F(M)	19M	×	×	**16.73**
TDNN-F(L)	21M	×	×	17.24

In the first set of experiments, we intend to examine the two preprocessing approaches mentioned in Section 2.1, namely data cleaning and lexicon augmentation, whose corresponding results are shown in Table 2. As can be seen from Table 2, when the data cleansing approach is applied, our DNN-HMM system can yield a relative SER reduction of 10.7% compared to the baseline system (TDNN-F) provided by the organizer of the FSR-2020 challenge. By comparison, when data cleaning is further paired with the lexicon augmentation, only a moderate improvement can be obtained. We hence conjecture that data cleansing is an indispensable component in the preprocessing stage for acoustic modeling. In addition, we also conduct a model ablation study to check whether stacking more layers to form a deeper neural network for acoustic modeling can lead to better performance. We find that it is not always the case when we add more layers to form a deeper neural network for acoustic modeling. On top of the best result drawn from Table 2, we will use TDNN-F(M) as our default acoustic model for the following experiments.

3.3 Experiments on Data Augmentation

Table 3. SER (%) results on the pilot-test dataset with disparate data augmentation methods.

Noise Injection	SpecAugment	Training epochs	SER (%)
-	-	6	16.73
×	-	6	16.36
×	×	6	15.22
×	×	12	**14.68**

In addition to the speed perturbation technique that is employed in the baseline setup, we also consider the use of another two data augmentation methods in building our ASR systems, viz. noise injection and spectrogram augmentation (*cf*. Section 2.2). Notably, we used our default acoustic model for this purpose, while the corresponding results are depicted in Table 3. It is evident from these results that the inclusion and combination of these two data augmentation methods can considerably boost the ASR performance. Among other things, due to the under-fitting problem incurred by SpecAugment (Park *et al.,* 2019), a further increase of the training epochs from 6 to 12 leads to the best relative SER reduction of 12.2% in relation to the baseline system.

3.4 Experiments on Acoustic and Language Modeling

Table 4. SER (%) results on the pilot-test dataset with different combinations of the acoustic and language models.

Acoustic Model	SER (%)	
	pilot-test	noisy-pilot-test
Baseline (TDNNF)	14.68	26.49
+ Four-gram LM	**13.47**	25.13
E2E-TDNNF	16.09	24.08
+ Four-gram LM	14.58	**21.71**

To make our pilot-test dataset consistent with the final test dataset which will be corrupted by varying noise sources, we also inject different types of noise, with SNR levels ranging from 5 to 20 dB, into the pilot-test dataset to form a noisy version of the training dataset that can be additionally exploited for training the acoustic model. Furthermore, we also evaluate two kinds of acoustic models, viz. baseline (TDNNF) and E2E-TDNNF (*cf*. Section 2.4), with the same model architecture and the best data augmentation setting made in Table 3. The corresponding results are shown in Table 4, from which an interesting phenomenon can be observed: when with the clean-condition testing setup, E2E-TDNNF performs worse than TDNNF that is based on the regular training configuration. On the contrary, in the noisy-condition testing setup, the E2E-TDNNF demonstrates superior noise-robustness performance.

On the other direction, when we adopt the four-gram language model (in replace of the trigram language model) for the ASR system to decode syllable sequences, a significant SER reduction can be obtained (*cf*. the last two rows of Table 4). As a side note, we also put effort into training variants of the recurrent neural network (RNN)-based language model for second-pass lattice rescoring (Xu *et al.,* 2018) (Wang *et al.,* 2019) (Chiu & Chen, 2021). Their performance, however, was not as good as expected, and we thus omit the details on the lattice rescoring experiments. From now on, unless otherwise stated, all variants of our ASR system discussed in the following experiments will use the four-gram language model for the decoding of syllable sequences.

3.5 Front-end Speech Enhancement

Table 5. SI-SNR (dB) results on the noisy pilot-test dataset with the front-end SE component.

SE model	SI-SNR (dB)	
	dev	noisy-pilot-test
No processing	6.82	7.26
Conv-TasNet (Luo & Mesgarani, 2019)	**16.43**	**17.37**

Table 6. SER (%) results on the noisy pilot-test dataset with the front-end SE component.

Acoustic Model	SER (%)	
	noisy-pilot-test	
	no process	enhanced
E2E-TDNNF	21.71	20.93
E2E-TDNNF-ENH	**20.76**	**19.57**

In an attempt to confirm the noise-robustness ability of our ASR system, we conduct a set of experiments with the front-end SE method, viz. Conv-TasNet, which aims at noise suppression. To train Conv-TasNet, we randomly set aside a portion of the training dataset as the development set, and followed the best criterion for training Conv-TasNet that was suggested by (Luo & Mesgarani, 2019), viz. minimization of the negative SI-SNR loss. Note here that the so-called permutation invariant training (PIT) was not employed. As can be seen from Table 5, when Conv-TasNet is applied, the SI-SNR results on both the development and pilot-test datasets can be considerably improved. Meanwhile, the SER performance of our ASR system can be substantially promoted (*cf*. the first row of Table 6).

To further enhance the acoustic modeling of our ASR system, we additionally augment the training dataset with a copy of the noisy training utterances which was processed by Conv-TasNet. As such, the augmented training dataset includes the original training data released by the organizer, its noise-contaminated counterpart and its noise-contaminated counterpart further enhanced by Conv-TasNet. We refer to this acoustic model as “E2E-TDNNF-ENH” in contrast to the original one (viz. E2E-TDNNF); E2E-TDNNF-ENH is created to simulate the effect of retraining of the acoustic model on enhanced speech signals. Inspection of the last row of Table 6, we can see that E2E-TDNNF-ENH can bring about a substantial SER improvement in comparison to E2E-TDNNF when the test utterances to be fed into the acoustic model were also enhanced by Conv-TasNet a priori. Nevertheless, in our experiments we spot-checked a few utterances of the pilot-test dataset that were enhanced by Conv-TasNet, and found that Conv-TasNet sporadically eliminated the speech portions of test utterances (viz. most of the speech portions became silent), which probably would lead to deteriorated ASR performance on the unseen test utterances. To secure a reliable performance level of our ASR system on the final test dataset, Conv-TasNet was merely used to obtain an enhanced-copy of the noisy training data, with the purpose of data augmentation for training E2E-TDNNF-ENH. Namely, Conv-TasNet will not be used to enhance the utterances of the final test dataset.

3.6 Transfer Learning and Semi-supervised Learning

Table 7. SER (%) results on the pilot-test dataset with transfer learning and semi-supervised learning.

Acoustic Model	SER (%)
	noisy-pilot-test
TDNNF-NER	25.15
CNN-TDNNF-NER	23.69
CNN-TDNNF-NER (with semi-supervised training)	**22.68**

In this paper, we also seek to capitalize on more training techniques for acoustic modeling in the context of Taiwanese ASR. To this end, we adopt the strategy proposed in (Lo & Chen, 2019), which in essence involved two techniques: transfer learning (Ghahremani *et al.,* 2017) and semi-supervised training (Manohar *et al.,* 2018). In implementation, we first used the weight transfer strategy (Ghahremani *et al.,* 2017) to train an acoustic model with parts of its model parameters transferred from a source model that were well-trained on the NER dataset (Liao *et al.,* 2017) beforehand. On a separate front, we also attempt to make use the label-agnostic final test dataset (viz. the corresponding reference transcripts of the final test dataset were not provided) to perform semi-supervised training of the acoustic model. In implementation, the recipe proposed in (Manohar *et al.,* 2018) was adopted, which used the entire lattice pertaining to each unlabeled training utterance as the supervision. The corresponding results are shown in Table 7, from which several observations can be drawn. First, when the TDNNF-based acoustic model was trained with transfer learning (denoted by TDNNF-NER), the SER result is slightly degraded compared to the result (25.13%) listed in the second row of Table 4. Second, if the acoustic model was built on top of the CNN-TDNNF structure, transfer learning can offer a considerable SER improvement on the noisy pilot-test dataset (*cf*. the second row of Table 7). It should be noted here that the CNN-TDNNF-based acoustic model, merely trained on the 50-hour training dataset offered by the FSR-2020 challenge, yields SER results significantly lower than 25.13%. This to some extent reveal that as opposed to TDNNF, CNN-TDNNF requires a larger amount of training dataset to achieve better ASR performance. In addition, when the label-agnostic final test dataset was additionally exploited to fine-tune the acoustic model, the performance of our ASR system on the pilot-test dataset can be boosted by a significant margin.

3.7 System Combination

Table 8. SER (%) results achieved by our two system-ensemble approaches on the noisy pilot-test dataset and final-test dataset.

Combined systems	SER (%)	
	noisy-pilot-test	final-test
+ CNN-TDNNF-NER + E2E-TDNNF + E2E-TDNNF-ENH	19.33	13.60
+ CNN-TDNNF-NER (Semi-supervised) + E2E-TDNNF + E2E-TDNNF-ENH	**19.10**	**13.10**

In the last set of experiments, we report on the results of our ASR systems submitted to FSR-2020 challenge, which were built based on two system-ensemble approaches that make combinations of different ASR systems previously evaluated in the above subsections. To be specific, we first performed lattice combination to merge all of the word lattices generated by different ASR systems into a single one with equal prior weights. Then, minimum Bayes-risk (MBR) decoding was conducted to obtain the ultimate ASR output for each test utterance. Here we combine the first three of the best systems according to their performance on the noisy pilot-test dataset. Table 8 shows the SER results of our two system-ensemble approaches on the noisy pilot-test dataset and final-test dataset. It is clear that these two system-ensemble approaches can substantially improve the ASR performance of our system on the pilot-test dataset.

3.8 Summary of the Experiments

Finally, we summarize the SER results of the participating teams on the final test dataset of Track 3 in the FSR-2020 challenge. Figures 2 and 3 show the SER evaluations of Track 3 with and without consideration of the correctness of tone transcription, respectively. Note here that although each team could submit two disparate results for evaluations, we only list the best result of each team here for brevity. Our ASR system has achieved the best performance among all participating teams for the two evaluation settings.

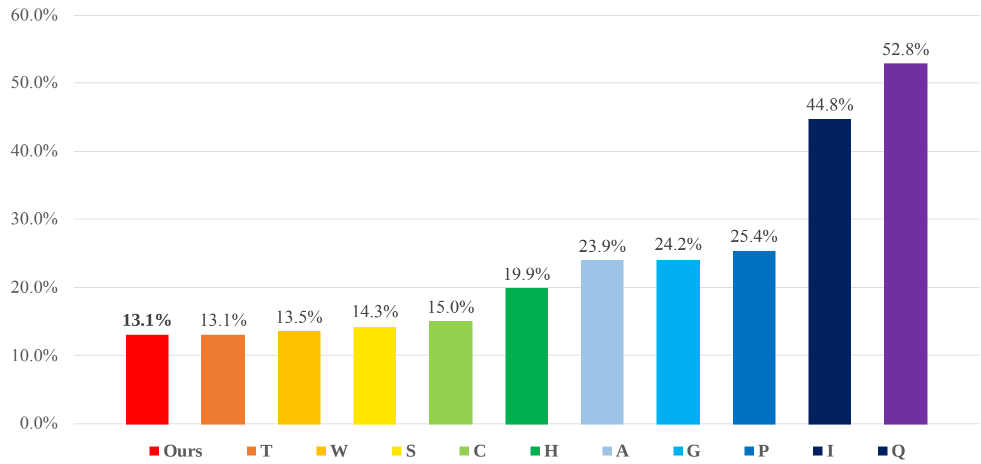

Figure 2. SER (%) results of all participating teams submitted to Track 3 in the FSR-2020 challenge (with consideration of the correctness tone transcription).

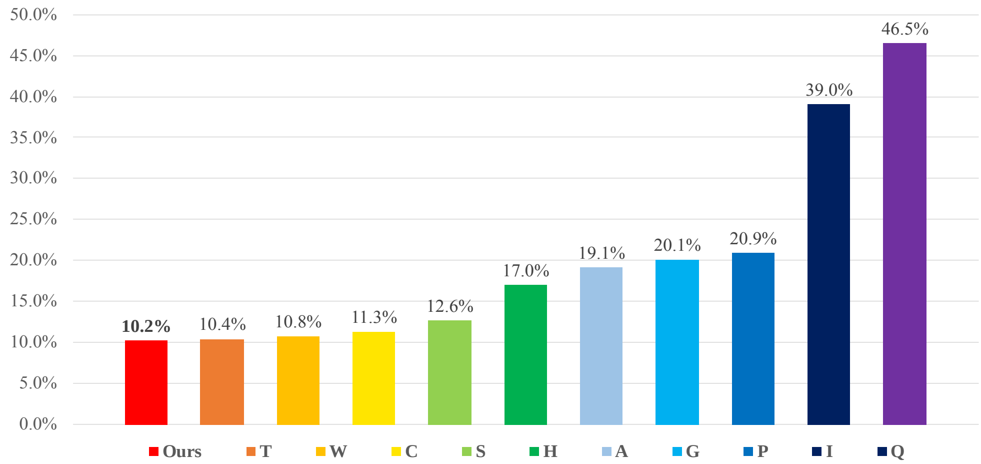

Figure 3. SER (%) results of all participating teams submitted to Track 3 in the FSR-2020 challenge (without consideration of the correctness tone transcription).

4. Conclusion and Future Work

In this paper, we have presented the modeling details of the NTNU ASR system that participated in the FSR-2020 Challenge. Through a series of experimental evaluation, the promising effectiveness of the joint use of data cleansing, data augmentation, front-end approach, transfer learning and semi-supervised learning methods for Taiwanese speech ASR has been confirmed. As to the future work, we plan to investigate more sophisticated end-to-end approaches for use in acoustic modeling of the Taiwanese language, as well as to apply our modeling strategy to other resource-poor ASR tasks.

References

Chiu, S. -H. & Chen, B. (2021). Innovative BERT-based reranking language models for speech recognition. In *Proc. of SLT 2021*, 266-271. https://doi.org/10.1109/SLT48900.2021.9383557

Dean, D. B., Sridharan, S., Vogt, R. J. & Mason, M. W. (2010). The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms. In *Proc. of Interspeech 2010*, 3110-3113.

Gale, W. A. (1995). Good-Turing smoothing without tears. *Journal of Quantitative Linguistics*, *2*, 217-237.

Ghahremani, P., Manohar, V., Hadian, H., Povey, D. & Khudanpur, S. (2017). Investigation of transfer learning for ASR using LF-MMI trained neural networks. In *Proc. of ASRU 2017*, 279-286. https://doi.org/10.1109/ASRU.2017.8268947

Hadian, H., Sameti, H., Povey, D. & Khudanpur, S. (2018). End-to-end speech recognition using lattice-free MMI. In *Proc. of Interspeech 2018*, 12-16. https://doi.org/10.21437/Interspeech.2018-1423

Jaitly, N. & Hinton, G. E. (2013). Vocal tract length perturbation (VTLP) improves speech recognition. In *Proc. of ICML 2013*.

Kinoshita, K., Ochiai, T., Delcroix, M. & Nakatani, T. (2020). Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In *Proc. of ICASSP 2020*, 7009-7013. https://doi.org/10.1109/ICASSP40776.2020.9053266

Ko, T., Peddinti, V., Povey, D. & Khudanpur, S. (2015). Audio augmentation for speech recog-nition. In *Proc. of Interspeech 2015*, 3586-3589.

Liao, Y. -F., Chang, Y. -H. S., Wang, S. -Y., Chen, J. -W., Wang, S. -M. & Wang, J.-H. (2017). A progress report of the Taiwan Mandarin Radio Speech Corpus Project. In *Proc. of O-COCOSDA 2017*. https://doi.org/10.1109/ICSDA.2017.8384450

Liao, Y.-F., Chang, C.-Y., Tiun, H.-K.., Su, H.-L., Khoo, H.-L., Tsay, J. S., Tan, L.-K., Kang, P., Thiann, T.-g., Iunn, U.-G., Yang, J.-H.,…Liang, C.-N. (2020). Formosa Speech Recognition Challenge 2020 and Taiwanese Across Taiwan Corpus. In *Proceedings of 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA ‘20)*, 65-70. https://doi.org/10.1109/O-COCOSDA50338.2020.9295019

Lo, T. -H. & Chen, B. (2019). Semi-supervised training of acoustic models leveraging knowledge transferred from out-of-domain data. In *Proc. of APSIPA ASC 2019*, 1400-1404. https://doi.org/10.1109/APSIPAASC47483.2019.9023040

Luo, Y. & Mesgarani, N. (2019). Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(8), 1256-1266. https://doi.org/10.1109/TASLP.2019.2915167

Manohar, V., Hadian, H., Povey, D. & Khudanpur, S. (2018). Semi-supervised training of acoustic models using lattice-free MMI. In *Proc. of ICASSP 2018*. https://doi.org/10.1109/ICASSP.2018.8462331

Ney, H. & Essen, U. (1991). On smoothing techniques for bigram-based natural language modelling. In *Proc. of ICASSP 1991*, 825-828. https://doi.org/10.1109/ICASSP.1991.150464

Park, D. S., Chan, W., Zhang, Y., Chiu, C. -C., Zoph, B., Cubuk, E. D., & Le, Q. V. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proc. of Interspeech 2019*, 2613- 2617. https://doi.org/10.21437/Interspeech.2019-2680

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., & Khudanpur, S. (2018). Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Proc. of Interspeech 2018*, 3743-3747. https://doi.org/ 10.21437/Interspeech.2018-1417

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., &Vesely, K. (2011). The Kaldi speech recognition toolkit. In *Proc. of ASRU 2011*.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., & Khudanpur, S. (2016). Purely sequence-trained neural networks for ASR based on lattice-free mmi. In *Proc. of Interspeech 2016*, 2751-2755. https://doi.org/10.21437/Interspeech.2016-595

Saki, F. & Kehtarnavaz, N. (2016). Automatic switching between noise classification and speech enhancement for hearing aid devices. In *Proc. of EMBC 2016*, 736-739. https://doi.org/10.1109/EMBC.2016.7590807

Saki, F., Sehgal, A., Panahi, I. & Kehtarnavaz, N. (2016). Smart phone-based real-time classification of noise signals using subband features and random forest classifier. In *Proc. of ICASSP 2016*, 2204-2208. https://doi.org/10.1109/ICASSP.2016.7472068

Snyder, D., Chen, G. & Povey, D. (2015). MUSAN: A music, speech, and noise corpus. In arXiv preprint arXiv:1510.08484

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proc. fo ICASSP 2002*, 901-904.

Taiwanese Hokkien. (2021, July 8). In Wikipedia. https://en.wikipedia.org/wiki/Taiwanese_Hokkien

Thiemann, J., Ito, N. & Vincent, E. (2013). The diverse environments multichannel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings.

The Journal of the Acoustical Society of America, *133*(5), 3591. https://doi.org/10.1121/1.4806631

Wang, C., Li, M. & Smola, A. J. (2019). Language models with transformers. In arXiv preprint arXiv:1904.09408

Xu, H., Li, K., Wang, Y., Wang, J., Kang, S., Chen, X., Povey, D., & Khudanpur, S. (2018). Neural network language modeling with letter-based features and importance sampling. In *Proc. of ICASSP 2018*, 6109-6113. https://doi.org/10.1109/ICASSP.2018.8461704

Computational Linguistics and Chinese Language Processing
Vol. 26, No. 1, June 2021, pp. 17-32

NSYSU-MITLab 團隊於福爾摩沙語音辨識競賽 2020 之語音辨識系統

NSYSU-MITLab Speech Recognition System for Formosa Speech Recognition Challenge 2020

林洪邦*、陳嘉平*

Hung-Pang Lin and Chia-Ping Chen

摘要

本論文中，我們描述了 NSYSU-MITLab 團隊在福爾摩沙語音辨識競賽 2020 (Formosa Speech Recognition Challenge 2020, FSR-2020) 中所實作的系統。我們使用多頭注意力機制 (Multi-head Attention) 所構成的 Transformer 架構建立了端到端的語音辨識系統，並且結合了連續性時序分類 (Connectionist Temporal Classification, CTC) 共同進行端到端的訓練以及解碼。我們也嘗試將編碼器更改為結合卷積神經網路 (Convolutional neural network, CNN) 與多頭注意力機制的 Conformer 架構。同時我們也建立了深度神經網路結合隱藏式馬可夫模型 (Deep Neural Network-Hidden Markov Model, DNN-HMM) ，其中我們以時間限制自注意力機制 (Time-Restricted Self-Attention, TRSA) 及分解時延神經網路 (Factorized Time Delay Neural Network, TDNN-F) 建立深度神經網路的部分。最終我們在台文漢字任務上得到最佳的字元錯誤率 (Character Error Rate, CER) 為 43.4% 以及在台羅拼音任務上取得最佳的音節錯誤率 (Syllable Error Rate, SER) 25.4%。

Abstract

In this paper, we describe the system team NSYSU-MITLab implemented for Formosa Speech Recognition Challenge 2020. We use the Transformer architecture

* 國立中山大學資訊工程學系
Department of Computer Science and Information Engineering National Sun Yat-sen University
E-mail: m083040013@g-mail.nsysu.edu.tw; cpchen@mail.cse.nsysu.edu.tw

composed of Multi-head Attention to construct an end-to-end speech recognition system and combine it with Connectionist Temporal Classification (CTC) for end-to-end training and decoding. We have also built a deep neural network combined with a hidden Markov model (DNN-HMM). We use Time-Restricted Self-Attention and Factorized Time Delay Neural Network (TDNN-F) for the deep neural network in DNN-HMM. The best performance we have achieved with the proposed methods is the character error rate of 45.5% for Taiwan Southern Min Recommended Characters (台文漢字) task and syllable error rate 25.4% for Taiwan Minnanyu Luomazi Pinyin (台羅拼音) task.

關鍵詞：自動語音辨識、Transformer、Conformer、連續性時序分類、聲學模型

Keywords: Automatic Speech Recognition、Transformer、Conformer、Connectionist Temporal Classification、Acoustic Model

1. 緒論 (Introduction)

近年來由於硬體效能的提昇以及深度學習 (Deep Learning) 理論的蓬勃發展，人工智慧成為了熱門的研究議題，無論在圖像、文字、語音等應用上皆可看到其蹤影。自動語音辨識 (Automatic Speech Recognition, ASR) 是深度學習中一個重要的應用， 在過往經常使用隱藏式馬可夫模型-高斯混合模型 (Hidden Markov Model-Gaussian Mixture Model, HMM-GMM) 來建立語音辨識模型，之後隨著深度學習的發展，出現了結合深度神經網路 (Deep Neural Network) 的 DNN-HMM。時延神經網路 (Time Delay Neural Network, TDNN) 以及長短期記憶 (Long Short-Term Memory, LSTM) 所構成的神經網路應用在 DNN-HMM 上獲得了不錯的成效(Peddinti *et al.,* 2018)，而時間限制自注意力機制 (Time-Restricted Self-Attention, TRSA) 也被證實可以用來代替網路中的 TDNN 或 LSTM (Povey *et al.,* 2018)。近年來自動語音辨識中端到端 (End-to-End) 方法成為了另一個熱門的研究項目，端到端的語音辨識系統能夠將聲音訊號輸入至一個模型便可直接輸出對應的文字序列。連續時序性分類 (Connectionist Temporal Classification, CTC) (Graves *et al.,* 2006)以及基於多頭注意力機制 (Multi-head Attention) 的 Transformer 架構 (Vaswani *et al.,* 2017) (Dong *et al.,* 2018)皆可做為端到端語音辨識的模型，其中又出現了結合兩者的目標函數共同進行訓練混合模型(Karita *et al.,* 2019)。

現今的語音辨識系統在英文以及中文等資源豐富的語言上已經能達到很好的辨識效果了，但是對於台語等資源相對較稀少的語言卻尚未有足夠的研究。我們參加了台語語音辨識競賽 FSR-2020 (Liao *et al.,* 2020)，並針對台文漢字任務建立了端到端的台語語音辨識模型，以及針對台羅拼音數字調任務建立了 DNN-HMM 以及端到端的台語語音辨識模型。本文分為四個部分：第一部分為緒論；第二部分為研究方法，介紹端到端的語音辨識模型架構以及 DNN-HMM 的模型架構；第三部分為實驗，介紹資料集、實驗設置以及實驗結果；第四部分為結論。

2. 研究方法 (Research Methods)

2.1 端到端模型架構 (End-to-End Model Architecture)

2.1.1 Transformer編碼器 (Transformer Encoder)

我們使用的端到端模型架構由 Transformer 的編碼器以及解碼器所構成，並且利用編碼器的輸出計算連續時序性分類 (Connectionist Temporal Classification, CTC)，模型架構如圖 1 所示。我們使用的聲學特徵是 80 維的 Fbank (Filter bank) 加上 3 維的音調 (Pitch)，首先我們會先以兩層 att 個卷積核 (kernel)、步長 (stride size) 2 、卷積核大小為 3 的卷積神經網路 (Convolutional neural network, CNN) 降低輸入特徵的維度，得到新的特徵 $X^{sub} \in R^{seq \times att}$ ，其中 att 為注意力機制的特徵維度大小。第 i 層編碼器計算流程如下：

$$X_0 = X^{sub} + PE \tag{1}$$

$$X_i' = \text{Layernorm}\big(X_i + \text{MHA}(X_i, X_i, X_i)\big) \tag{2}$$

$$X_{i+1} = \text{Layernorm}\big(X_i' + \text{FF}(X_i')\big) \tag{3}$$

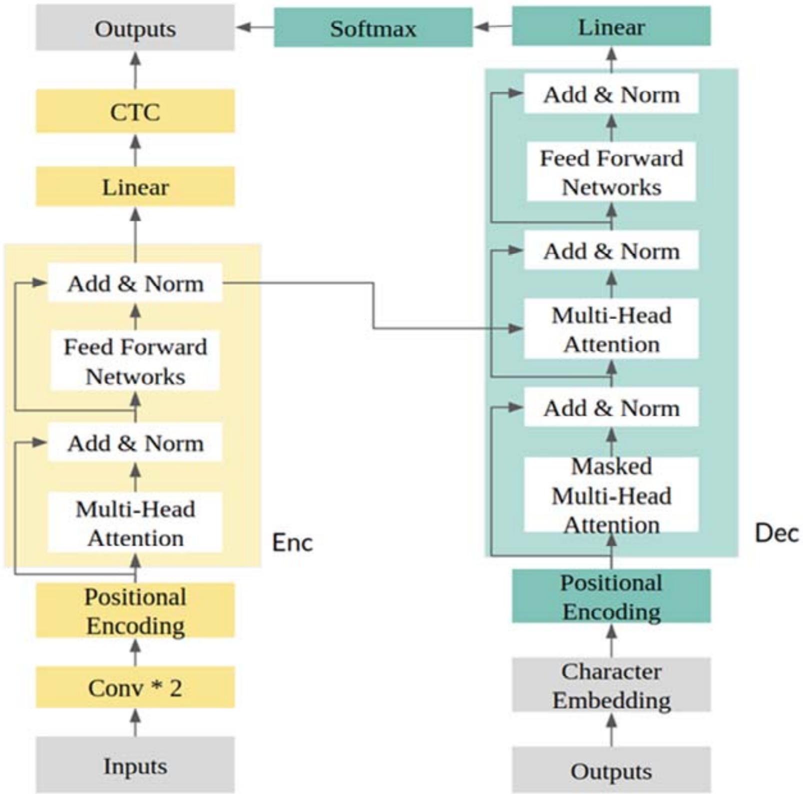

圖 1. Transformer-based 模型架構
[Figure 1. Transformer-based model architecture]

由於多頭注意力機制無法取得輸入序列的前後位置資訊，因此輸入編碼器的特徵要再額外加上位置編碼 PE (Positional Encoding) (Vaswani *et al.*, 2017)。多頭注意力機制 MHA 的計算方式如下：

$$\text{attention}(Q,K,V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{att}}\right)V \tag{4}$$

$$H_h = \text{attention}\left(QW_h^q, KW_h^k, VW_h^v\right) \tag{5}$$

$$\text{MHA}(Q,K,V) = [H_1, H_2, \ldots, H_{head}]W^{head} \tag{6}$$

其中 $K,V \in R^{k\times att}$ 以及 $Q \in R^{q\times att}$ 是 MHA 層的輸入，$W_h^q, W_h^k, W_h^v \in R^{att\times(att/head)}$ 以及 $W^{head} \in R^{att\times att}$ 是可以學習的參數，att 為注意力機制的特徵維度大小。head 個自注意力機制計算的結果 H_h 將串接起來，並與 MHA 的輸入進行殘差連結 (Residual connect) 後再通過層標準化 (Layer Normalization) (Ba *et al.*, 2016)得到最終的輸出。公式(3) 中的 FF 則是前饋神經網路 (Feed-Forward Neural Network)，由兩層全連接層以及 Rectified Linear Unit (ReLU) 激活函數所構成。在經過數層編碼器的運算後，得到最終的編碼器輸出 $X_e \in R^{seq\times att}$ 。

2.1.2 Transformer 解碼器 (Transformer Decoder)

得到 Transformer 編碼器的輸出 X_e 後，Transformer 解碼器會以 X_e 以及前一個時刻的字元序列 $Y[1{:}u]$ 計算下一個輸出字元 $Y[u+1]$ 的機率。第 j 層解碼器的計算流程如下：

$$E = \text{Embed}(Y[1{:}u]) \tag{7}$$

$$Z_0 = E + PE \tag{8}$$

$$Z_j' = \text{Layernorm}\left(Z_j + \text{MHA}\left(Z_j, Z_j, Z_j\right)\right) \tag{9}$$

$$Z_j'' = \text{Layernorm}\left(Z_j' + \text{MHA}\left(Z_j', X_e, X_e\right)\right) \tag{10}$$

$$Z_{j+1} = \text{Layernorm}\left(Z_j'' + \text{FF}\left(Z_j''\right)\right) \tag{11}$$

其中公式(7) 會將字元序列轉換為詞嵌入 (Word Embedding) $E \in R^{u\times att}$，加上位置編碼後做為第一層解碼器的輸入 Z_0。下一個輸出字元 $Y[u+1]$ 的後驗機率計算方式如下：

$$[p_{s2s}(Y[2]|Y[1],X_e), \ldots, p_{s2s}(Y[u+1]|Y[1{:}u],X_e)] = \text{softmax}(Z_d W^{att} + b^{att}) \tag{12}$$

$$p_{s2s}(Y|X_e) = \prod_u p_{s2s}(Y[u+1]|Y[1{:}u],X_e) \tag{13}$$

其中 Z_d 為最後一層解碼器的輸出， $W^{att} \in R^{att\times token}, b^{att} \in R^{token}$ 則是可以學習的參數，在經過 softmax 的運算後可以得到每個可能字元的輸出機率。

2.1.3 Conformer編碼器 (Conformer Encoder)

除了 Transformer 編碼器外，我們也嘗試使用結合 CNN 與 Transformer 的 Conformer 編碼

器 (Gulati *et al.*, 2020)，透過卷積模組 (Convolution Module) 以及 多頭注意力機制，Conformer 能更好的取得局部以及全局的訊息，模型架構如圖 2 所示。第 i 層 Conformer 編碼器的計算流程如下：

$$X_i' = X_i + \frac{1}{2}\mathrm{FF}(\mathrm{Layernorm}(X_i)) \tag{14}$$

$$X_i'' = X_i' + \mathrm{MHA}(\mathrm{Layernorm}(X_i')) \tag{15}$$

$$X_i''' = X_i'' + \mathrm{Conv}(X_i'') \tag{16}$$

$$X_{i+1} = \mathrm{Layernorm}\left(X_i''' + \frac{1}{2}\mathrm{FF}(\mathrm{Layernorm}(X_i'''))\right) \tag{17}$$

Conformer 使用馬卡龍網路 (Macaron-Net) (Lu *et al.*, 2019)的架構，將兩個前饋神經網路置於編碼器的頭尾，並且將前饋神經網路中的 ReLU 激活函數替換為 Swish 激活函數 (Ramachandran *et al.*, 2017)。另外 Conformer 採用源於 Transformer-XL 的相對位置編碼 (Relative Positional Encoding) (Dai et al. 2019)，相較於一般的位置編碼，相對位置編碼在面對不同長度的輸入能取得較好的位置資訊。在多頭注意力機制之後，Conformer 加入了卷積模組，卷積模組從單位卷積層 (Pointwise Convolution) 以及門控線性單元 (Gated Linear Unit, GLU) (Dauphin *et al.*, 2017)開始，接著會再經過深度卷積層 (Depthwise Convolution)、批量標準化 (Batch Normalization)、Swish 激活函數及單位卷積層。

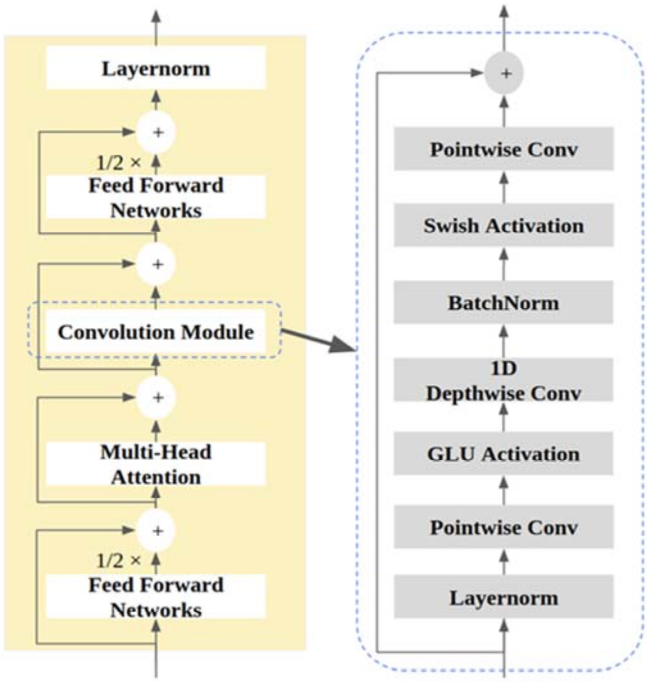

圖 2. Conformer 模型架構
[Figure 2. Conformer model architecture]

2.1.4 混合訓練及解碼 (Hybrid Training and Decoding)

除了解碼器的輸出外，我們也會利用編碼器的輸出計算連續時序性分類 (CTC)，其計算方式如下：

$$C = \mathrm{softmax}(X_e W^{ctc} + b^{ctc}) \tag{18}$$

$$p(\pi|X_e) = \prod_{t=1}^{seq} C\big[t, \pi[t]\big] \tag{19}$$

$$p_{ctc}(Y|X_e) = \sum_{\pi\in\beta^{-1}(Y)} p(\pi|X_e) \tag{20}$$

其中 $W^{ctc} \in R^{att\times token}, b^{ctc} \in R^{token}$ 是可以學習的參數。由於輸入序列的長度一般會大於輸出序列的長度，因此輸出會有多種可能的組合，這些組合記為π，公式(19)中 $C\big[t, \pi[t]\big]$ 代表 X_e 的第 t 個幀輸出 $\pi[t]$ 的機率。$\beta(\pi)$ 是一個多對一的函式，會將 π 中冗餘的字元刪除，例如 $\beta(a\emptyset aabb) = aab$ 。公式(20)中 $\beta^{-1}(Y) = \{\pi|Y = \beta(\pi)\}$ 代表所有可以形成文字序列 Y 的組合。最終我們將混合解碼器以及 CTC 的輸出，得到損失函數：

$$L_{mtl} = -\alpha \log p_{s2s}(Y|X_e) - (1-\alpha)\log p_{ctc}(Y|X_e) \tag{21}$$

其中 α 為超參數，控制解碼器及 CTC 在損失函數中所佔的比例。在解碼階段時，除了解碼器以及 CTC 的輸出外，我們額外加入語言模型共同進行解碼，解碼方式如下：

$$\hat{Y} = \arg\max_{Y\in y^*}\{\lambda \log p_{s2s}(Y|X_e) + (1-\lambda)\log p_{ctc}(Y|X_e) + \gamma \log p_{lm}(Y)\} \tag{22}$$

其中y^* 代表輸出的候選字集合，$p_{lm}(Y)$ 則代表語言模型派給候選序列 Y 的機率。λ 及 γ 為超參數，分別控制在解碼階段時解碼器、CTC 以及語言模型所佔的比例。

2.2 DNN-HMM

2.2.1 分解時延神經網路 (Factorized Time Delay Neural Network)

除了端到端的模型架構外，我們也嘗試針對台羅拼音任務建立 DNN-HMM 的語音辨識模型。在聲學模型中，我們使用到了分解時延神經網路 (Factorized Time Delay Neural Network, TDNN-F) (Povey *et al.*, 2018)，此種模型架構拆解一般時延神經網路的權重矩陣，並在訓練過程中限制其中一個分解的權重矩陣保持正交，以維持訓練時的穩定度。假設 M 為參數矩陣，我們定義 $P \equiv MM^{\mathrm{T}}, Q \equiv P - I$，並希望透過參數更新最小化 $f = \mathrm{tr}(QQ^{\mathrm{T}})$ 使得 M 保持正交。參數更新的公式如下：

$$M \leftarrow M - \frac{1}{2\alpha^2}(MM^{\mathrm{T}} - \alpha^2 I)M \tag{23}$$

其中α 是縮放參數，設為$\sqrt{\mathrm{tr}(PP^{\mathrm{T}})/\mathrm{tr}(P)}$ 。圖 3 為 TDNN-F 的模型架構，1536 維的隱藏層會被拆解為 1536×160×1536 ，其中第一個權重矩陣會保持正交限制，之後再加上 ReLU 激活函數、批量標準化以及殘差連結 。

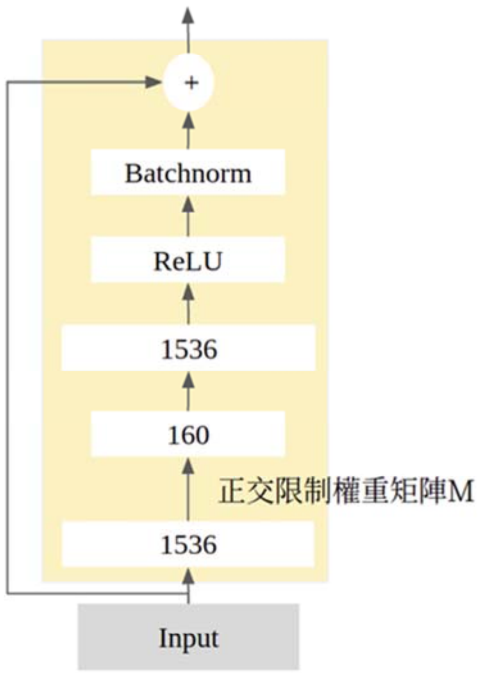

圖 3. TDNN-F 模型架構
[Figure 3. TDNN-F model architecture]

2.2.2 時間限制自注意力機制 (Time-Restricted Self-Attention)

除了 TDNN-F 外，我們也使用了時間限制自注意力機制 (Time-Restricted Self-Attention, TRSA) (Povey *et al.*, 2018)，此種特殊的自注意力機制在計算公式(4)時會限制模型的關注範圍，其計算方式如下：

$$\boldsymbol{y_t} = \sum_{\tau=t-L}^{t+R} c_t(\tau)\boldsymbol{v_t} \tag{24}$$

$$c_t(\tau) = \frac{\exp(\boldsymbol{q_t}\cdot\boldsymbol{k_\tau}/\sqrt{d_k})}{\sum_{\tau'=t-L}^{t+R} exp(\boldsymbol{q_t}\cdot\boldsymbol{k_{\tau'}}/\sqrt{d_k})} \tag{25}$$

其中L, R 代表自注意力機制所能關注的最大左右範圍，$\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}$ 則是輸入 TRSA 層的向量，d_k 為向量 $\boldsymbol{q}, \boldsymbol{k}$ 的維度大小。$\boldsymbol{y_t}$ 為向量 $\boldsymbol{v_t}$ 的加權和，其權重為輸入向量 $\boldsymbol{q_t}$ 與前後 $L + R + 1$ 個 $\boldsymbol{k}$ 計算內積後通過 softmax 正規化所得到，之後再通過 ReLU 激活函數以及批量標準化後得到最終的輸出。同樣的 TRSA 也可以進行多頭 (Multi-head) 運算，將數個自注意力機制的結果串接起來得到輸出。

2.2.3 模型架構 (Model Architecture)

我們所使用的三種聲學模型架構如圖 4 所示，輸入的聲學特徵為 40 維的 MFCC 和 3 維的音調 (pitch) 加上 100 維的 i-vectors。首先輸入會先經過一層大小為 816，拼接窗 (splicing window) 為 (-1,0,1) 的 TDNN，接著會通過 6 層 TRSA-Transformer 架構，與一般 Transformer 架構不同的是我們使用的自注意力機制為時間限制自注意力機制。其中 TRSA 層的輸入向量 $\boldsymbol{q}, \boldsymbol{k}$ 的維度大小設為 40，$\boldsymbol{v}$ 的維度大小設為 60，多頭運算的數量設為 12，時間限制的範圍 L, R 則分別設為 5 以及 2。除此之外，我們也嘗試在 TRSA 層及前饋神經網路之間加入一層 TDNN-F，以及嘗試使用兩個前饋神經網路的馬卡龍網路。其中 TDNN-F 隱藏層的維度大小為 $1536 \times 160 \times 1536$，前饋神經網路的維度大小則設為 1024。我們所使用的目標函數為 Lattice-free Maximum Mutual Information (LF-MMI) 搭配交叉熵 (Cross Entropy) 的輔助正規化訓練(Povey *et al.*, 2016)，LF-MMI 使得訓練最大化交互資訊 (Maximum Mutual Information, MMI) 時不需要事先產生詞圖 (Word Lattices)，讓訓練過程能更加快速。

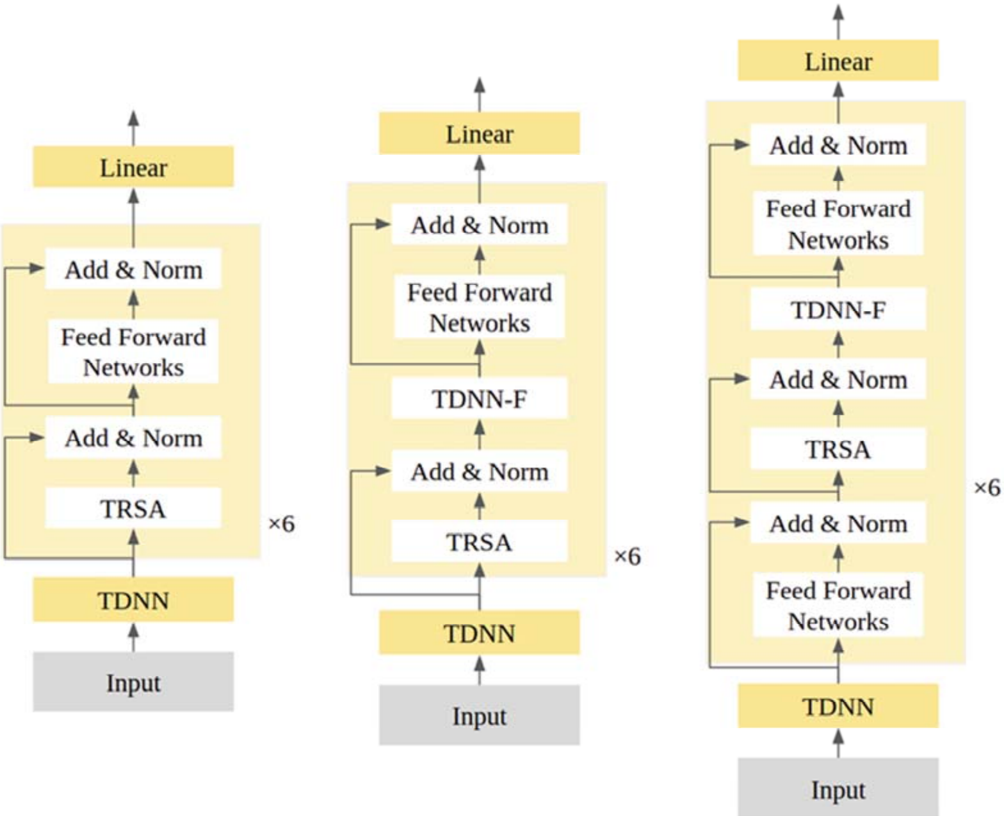

圖 4. 聲學模型架構
[Figure 4. Acoustic model architecture]

3. 實驗 (Experiments)

3.1 資料集 (Dataset)

3.1.1 Taiwanese across Taiwan (TAT) corpus

TAT-Vol1-train-lavalier 以及 TAT-Vol1-eval-lavalier 為經由競賽取得的資料集，錄音裝置為領夾式麥克風，取樣頻率為 16kHz，文本為台文漢字及台羅拼音。其中 TAT-Vol1-train-lavalier 共 80 位語者，總時長約 41.76 小時，共 23,104 筆音檔。TAT-Vol1-eval-lavalier 總時長約 4.78 小時，共 2,664 筆音檔。TAT-Vol1-train-lavalier 資料集的文本中，出現了台文漢字以及台羅拼音混用的情形。為了使同一種文本的標註一致，我們將台文漢字文本中出現非台文漢字的句子刪除，同樣的我們也將台羅拼音文本中出現非台羅拼音的句子刪除。此外我們也刪除了文本中的標點符號。

3.1.2 PTS_TW-train

PTS_TW-train 資料集為公視台語台「台灣記事簿」以及「台灣新眼界」的節目內容，文本為台文漢字，取樣頻率為 16kHz，共 95 筆音檔，總時長約 85.39 小時。TAT 以及 PTS_TW-train 資料集的數據如表 1。在 PTS_TW-train 資料集中，每句文本皆有其對應的時間標註，但是並非每筆時間標註皆精確無誤，且有部分音檔內容為中文，因此我們使用 CTC-Segmentation (Kürzinger *et al.*, 2020) 重新對齊資料集的文本及音檔。CTC-Segmentation 利用已經訓練完成且具有 CTC 輸出的模型對音檔解碼，並計算每一幀輸出欲對齊文本字元的機率，藉此取得文本的時間資訊。我們利用 CTC-Segmentation 對齊時的對數機率 (Log Probability) 選擇分數較高的音檔片段保留，最終我們根據兩種不同的對數機率閾值得到了資料集 PTS-1.5 以及 PTS-5.0，資料集的數據如表 2。

表 1. TAT 及 PTS_TW-train 資料集數據
[Table 1. Details of TAT and PTS_TW-train dataset]

資料集	音檔數	總時數	語者	文本
TAT-Vol1-train-lavalier	23,104	41.76	80	台文、台羅
TAT-Vol1-eval-lavalier	2,664	4.78	-	台文、台羅
PTS_TW-train	95	85.39	-	台文

表 2. PTS-1.5 及 PTS-5.0 資料集數據
[Table 2. Details of PTS-1.5 and PTS-5.0 dataset]

資料集	對數機率閾值	音檔數	總時數
PTS-1.5	-1.5	7,665	16.92
PTS-5.0	-5.0	23,363	52.71

3.2 實驗設置 (Experimental Setups)

我們使用兩種資料增強的方式，分別是針對音檔的變速擾動 (Speed-Perturbation) (Ko *et al.,* 2015)，額外產生語速 1.1 及 0.9 的音檔；以及針對頻譜圖進行遮蔽以及扭曲的 SpecAugment (Park *et al.,* 2019)，其中針對頻率遮蔽的大小設為 30、數量為 2，針對時間遮蔽的大小設為 40、數量為 2，針對時間扭曲的大小則設為 5。針對端到端模型，多頭注意力機制的 head 數設為 4， Conformer 的卷積模組中深度卷積層的卷積核大小設為 32，混合訓練中的超參數 α 設為 0.3，解碼階段的超參數 λ 以及 γ 分別設為 0.5 以及 0.7。我們使用 Adam 優化器 (Optimizer) (Kingma & Ba, 2014)，學習率的更新方式如 (Vaswani *et al.,* 2017) ，會在起始的 25,000 次參數更新中逐漸上升，隨後便線性下降。在解碼階段中，我們使用的語言模型為兩層 1024 個單元的長短期記憶 (Long Short-Term Memory, LSTM)，訓練語言模型的文本皆來自 TAT-Vol1-train-lavalier 及 PTS_TW-train 資料集。針對台羅拼音任務，除了使用端到端的語音辨識模型外，我們也以競賽方提供的詞典 (lexicon) 為基礎建立了 DNN-HMM 的語音辨識模型，詞典以台羅拼音的聲母以及韻母加上數字調作為音素。我們訓練了一個 monophone 的 HMM-GMM 模型以及五個 triphone 的 HMM-GMM 模型以取得訓練 DNN 時所需的狀態標籤 (State Label)，在解碼階段時則使用 tri-gram 語言模型。端到端模型使用開源套件 ESPnet (Watanabe *et al.,* 2018) 建立，並在 Nvidia GeForce GTX 1080 Ti GPU 上訓練 50 個 epochs，批量大小(Batch Size) 設為 32。DNN-HMM 模型則是使用開源套件 Kaldi (Povey *et al.,* 2011)建立，並在 Nvidia GeForce GTX 2080 Ti GPU 上訓練 12 個 epochs，批量大小設為 16。

3.3 實驗結果 (Results)

3.3.1 台文漢字實驗結果 (Hàn-jī Task Results)

針對台文漢字任務，我們使用端到端的語音辨識模型。首先，我們藉由調整 Transformer 及 Conformer 架構的層數、注意力機制的特徵維度大小以及前饋神經網路的維度大小，測試模型參數量對字元錯誤率的影響。表 3 為各種模型參數的組合及其對應的字元錯誤率，此實驗的訓練集使用 TAT-Vol1-train-lavalier，並在 TAT-Vole1-eval-lavalier 資料集上測試。由實驗結果可以看到，在相近的參數量大小時，使用 Conformer 編碼器能夠得到優於 Transformer 編碼器的結果。接著我們測試增加訓練集對模型效果的差異，我們將訓練集額外加入 PTS-1.5 以及 PTS-5.0，得到的實驗結果如表 4 所示。我們所使用的模型架構為表 3 中字元錯誤率最佳的模型，由實驗結果可以看出，加入額外的資料集進行訓練能顯著的提升模型的效能。圖 5 為各隊伍之排名及 CER。

表 3. 不同模型參數量的實驗結果
[Table 3. The results of different model parameters]

編碼器架構	Transformer	Conformer			
參數量(M)	29.5	9.9	29.6	30.7	45.4
編碼器層數	12	10	12	14	12
解碼器層數	6	4	6	4	6
注意力機制維度	256	128	256	256	256
前饋神經網路維度	2048	1024	1024	1024	2048
CER(%)	36.3	27.1	26.6	26.9	28.7

表 4. 不同訓練集的實驗結果
[Table 4. The results of the different training dataset]

訓練集	CER(%)	Final-Test CER(%)
TAT-Vol1-train-lavalier	26.6	-
TAT-Vol1-train-lavalier + PTS-1.5	21.2	48.0
TAT-Vol1-train-lavalier + PTS-5.0	19.1	43.4

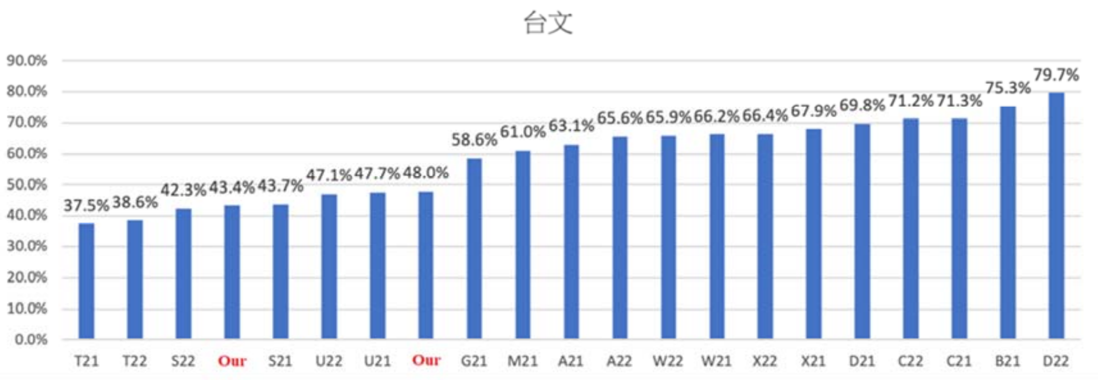

圖 5. 台文漢字競賽結果
[Figure 5. Hàn-jī challenge result]

3.3.2 台羅拼音實驗結果 (Tâi-lô Task Results)

針對端到端模型，我們比較 Transformer 編碼器以及 Conformer 編碼器的效果，以及使用不同輸出單位的差異。首先我們將每個字母及數字視為獨立的字元 (Char-based)，得到一共 37 個不同的字元；接著我們嘗試以音節為單位 (Syllable-based)，得到共 2212 個不同的音節;最後我們使用位元組對編碼 (Byte Pair Encoding, BPE) (Sennrich *et al.*, 2015)，將文本中最常出現的連續字元合併成新的 Subword，最後我們得到共 987 個不同的

Subword。表 5 為端到端模型的實驗結果，我們所使用的 Transformer 編碼器及 Conformer 編碼器的架構分別為表 3 中參數量為 29.5M 以及 29.6M 的模型。在同樣的模型架構下，使用 BPE 能得到優於其他輸出單位的結果。而在使用同樣的輸出單位時，Transformer 編碼器則取得了較佳的效果，其中以字元為單位的 Conformer 模型出現了無法收斂的情形。最終我們以使用 BPE 的 Transformer 模型在 Final-test 上取得了 25.4% 的音節錯誤率。針對 DNN-HMM，我們測試了圖 4 中三種聲學模型架構的差異，實驗結果如表 6 所示，當我們將 TDNN-F 加入 TRSA-Transformer 可以得到優於單純使用 TRSA-Transformer 的結果，而加入馬卡龍網路後又可以再進一步改進模型效能。最後我們也測試了加入 PTS 資料集的效果，由於該資料集的文本為台文漢字，因此我們利用 TAT-Vol1-train-lavalier 資料集取得共 2728 個台文漢字對應的台羅拼音，藉此取得 PTS 資料集的台羅拼音文本。實驗結果如表 7 所示，實驗中所使用的端到端模型及 DNN-HMM 模型分別為表 5 及表 6 中音節錯誤率最佳的模型，可以看出不論是端到端的語音辨識模型還是 DNN-HMM 的語音辨識模型皆隨著訓練資料量的增加而提升模型效能，其中又以端到端模型的提升效果更為明顯。圖 6 為各隊伍之排名及 SER。

表 5. 端到端模型實驗結果
[Table 5. The results of the end-to-end model]

編碼器架構	輸出單位	SER(%)	Final-Test SER(%)
Transformer	Char-based	20.8	-
	Syllable-based	20.9	-
	BPE	18.3	25.4
Conformer	Char-based	-	-
	Syllable-based	23.6	-
	BPE	19.3	-

表 6. DNN-HMM 實驗結果
[Table 6. The results of DNN-HMM]

模型架構	SER(%)	Final-Test SER(%)
TRSA-Transformer	16.3	-
TRSA-Transformer + TDNN-F	15.4	-
TRSA-Transformer + TDNN-F + 馬卡龍網路	15.3	27.1

表 7. 不同訓練集的實驗結果
[Table 7. The results of the different training dataset]

模型架構	訓練集	SER(%)
端到端語音辨識模型	TAT-Vol1-train-lavalier	18.3
	TAT-Vol1-train-lavalier + PTS-1.5	15.9
	TAT-Vol1-train-lavalier + PTS-5.0	14.3
DNN-HMM 語音辨識模型	TAT-Vol1-train-lavalier	15.3
	TAT-Vol1-train-lavalier + PTS-1.5	14.4
	TAT-Vol1-train-lavalier + PTS-5.0	13.7

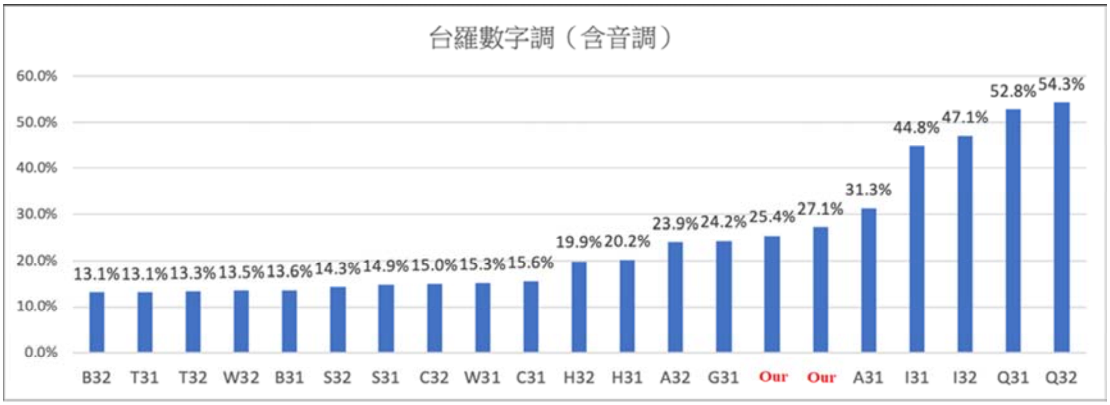

圖 6. 台羅拼音數字調競賽結果
[Figure 6. Tâi-lô challenge result]

4. 結論 (Conclusion)

我們描述了參加 FSR-2020 所使用的語音辨識系統，針對台文漢字的任務我們建立了端到端的語音辨識模型，其中編碼器使用 Conformer 架構，解碼器則為 Transformer 架構，同時也使用 CTC 損失函數共同進行訓練。我們也利用 CTC-Segmentation 對 PTS_TW-train 資料集進行前處理，以取得更多的訓練資料，最終我們取得最好的 CER 為 43.4%。針對台羅拼音任務的端到端模型則是使用 Transformer 編碼器，並且以 BPE 對文本編碼；除此之外也建立了 DNN-HMM 的語音辨識模型，聲學模型使用 TRSA-Transformer 架構結合 TDNN-F 以及馬卡龍網路。在台羅拼音任務上我們取得最好的 SER 為 25.4%。

參考文獻 (References)

Ba, J. L., Kiros, J. R. & Hinton, G. E. (2016). Layer normalization. In arXiv preprint arXiv:1607.06450

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. In arXiv preprint arXiv:1901.02860

Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D. (2017). Language modeling with gated convolutional networks. In *Proceedings of the 34th International conference on machine learning*, 933-941. https://doi.org/10.5555/3305381.3305478

Dong, L., Xu, S. & Xu, B. (2018). Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, 5884-5888. https://doi.org/10.1109/ICASSP.2018.8462506

Graves, A., Fernández, S., Gomez, F. & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recur-rent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, 369-376. https://doi.org/10.1145/1143844.1143891

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. In arXiv preprint arXiv:2005.08100

Karita, S., Soplin, N. E. Y., Watanabe, S., Delcroix, M., Ogawa, A. & Nakatani, T. (2019). Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration. In *Proc. of Interspeech 2019*, 1408-1412. https://doi.org/10.21437/Interspeech.2019-1938

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. In arXiv preprint arXiv:1412.6980

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Proceedings of Interspeech 2015*, 3586-3589.

Kürzinger, L., Winkelbauer, D., Li, L., Watzel, T., & Rigoll, G. (2020). CTC-segmentation of large corpora for German end-to-end speech recognition. In *Proceedings of 22nd International Conference on Speech and Computer (SPECOM 2020)*, 267-278. https://doi.org/10.1007/978-3-030-60276-5_27

Liao, Y.-F., Chang, C.-Y., Tiun, H.-K.., Su, H.-L., Khoo, H.-L., Tsay, J. S., Tan, L.-K., Kang, P., Thiann, T.-g., Iunn, U.-G., Yang, J.-H.,…Liang, C.-N. (2020). Formosa Speech Recognition Challenge 2020 and Taiwanese across Taiwan Corpus. In *Proceedings of 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA '20)*, 65-70. https://doi.org/10.1109/O-COCOSDA50338.2020.9295019

Lu, Y., Li, Z., He, D., Sun, Z., Dong, B., Qin, T., Wang, L., & Liu, T.-Y. (2019). Understanding and improving transformer from a multi-particle dynamic system point of view. In arXiv preprint arXiv:1906.02762

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. In arXiv preprint arXiv:1904.08779

Peddinti, V., Wang, Y., Povey, D. & Khudanpur, S. (2018). Low latency acoustic modeling using temporal convolution and lstms. *IEEE Signal Processing Letters*, *25*(3), 373-377. https://doi.org/10.1109/LSP.2017.2723507

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., & Khudanpur, S. (2018). Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Proceedings of Interspeech 2018*, 3743-3747. https://doi.org/10.21437/Interspeech.2018-1417

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The kaldi speech recognition toolkit. In *Proceedings of IEEE 2011 workshop on automatic speech recognition and understanding*.

Povey, D., Hadian, H., Ghahremani, P., Li, K. & Khudanpur, S. (2018). A time-restricted self-attention layer for asr. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, 5874-5878. https://doi.org/10.1109/ICASSP.2018.8462497

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., & Khudanpur, S. (2016). Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Proceedings of Interspeech 2016*, 2751-2755. https://doi.org/10.21437/Interspeech.2016-595

Ramachandran, P., Zoph, B. & Le, Q. V. (2017). Searching for activation functions. In arXiv preprint arXiv:1710.05941

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. In arXiv preprint arXiv:1508.07909

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. In arXiv preprint arXiv:1706.03762

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., & Ochiai, T. (2018). ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech 2018*, 2207-2211. https://doi.org/10.21437/INTERSPEECH.2018-1456

Computational Linguistics and Chinese Language Processing
Vol. 26, No. 1, June 2021, pp. 33-42

2020 福爾摩沙臺語語音辨識比賽之初步實驗

A Preliminary Study of Formosa Speech Recognition Challenge 2020 – Taiwanese ASR

余福浩*、盧克函*、王繹崴*、張維哲*、黃偉愷*、陳冠宇*

Fu-Hao Yu, Ke-Han Lu, Yi-Wei Wang,

Wei-Zhe Chang, Wei-Kai Huang and Kuan-Yu Chen

摘要

為研究當前深度學習語音辨識模型於臺文與臺羅拼音之語音辨識任務之成效，本研究使用 2020 福爾摩沙臺語語音辨識競賽(Formosa Speech Recognition Challenge 2020, FSR-2020)所提供之臺文語音語料庫(TAT-Vol1)以及公視臺語台訓練語料，並基於 ESPnet 與 Kaldi，比較數種模型架構、訓練方法與參數於臺語語音辨識之成效。最終，在 2020 福爾摩沙臺語語音辨識競賽裡，我們的系統在臺文辨識(Track2)中取得 66.1%的錯誤率，而在臺羅拼音辨識(Track3)方面，我們所得到的錯誤率為 28.6%。

Abstract

In order to study the effectiveness of the current deep learning-based speech recognition models in the speech recognition tasks of Taiwanese Southern Min Recommended Characters and Taiwan Minnanyu Luomazi Pinyin, this study uses the corpora provided by the 2020 Formosa Speech Recognition Challenge 2020 (FSR-2020) to evalutae some neural-based ASR systems by ESPnet and Kaldi toolkits. In the end, our system achieved a 66.1% error rate in the Taiwanese Southern Min Recommended Characters recognition (Track2), and the error rate we got in the Taiwan Minnanyu Luomazi Pinyin recognition (Track3) was 28.6%.

*國立臺灣科技大學資訊工程系
Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology
E-mail:{M10815004, M10915010, M10915036, M10915082, M10915012, kychen}@mail.ntust.edu.tw

關鍵詞：臺文，臺羅拼音，臺語語音辨識

Keywords: Taiwanese Southern Min Recommended Characters, Taiwan Minnanyu Luomazi Pinyin, Taiwanese ASR

1. 緒論 (Introduction)

臺語(Taiwanese)又稱為臺灣話、臺灣閩南語，是臺灣地區許多人的母語（本土語言），由於過去華語政策的影響，中文(Mandarin Chinese)逐漸取代各地區母語，包括臺語、客家語以及各族之原住民語，成為臺灣目前主流的溝通語言。然而，能夠流利使用傳統母語的人數逐年下滑，以聯合國教科文組織對語言提出的世代傳承指標(Intergenerational Language Transmission)來看，臺語已處於確定瀕危(Definitely Endangered)（指小孩不再於家中如同學習母語一樣學習此語言）至嚴重瀕危（只有祖父母或更老一輩的人會說此語言；儘管父母這一代人可能理解，但不會對小孩說）之間。因此，如果小孩不再學習及使用臺語，臺語極有可能在幾十年後就成為指標中滅絕(Extinct)的語言。

近年來由於深度學習的興起，在語音辨識(Speech Recognition)領域中，深度學習模型也已逐漸取代傳統機率式模型，在許多資料集上也早已超過傳統模型，獲得相當好的辨識成果。語音辨識資料集不僅蒐集不易，語料的標註(Labeling)構是耗時費工。臺語，因為不是主流語言，要找到熟悉臺語的人士進行語料標註，就顯得更加地不容易。因此要進行臺語語音辨識器的訓練，就得面對低資源(Low-resource)甚至是無資源的問題。為維護臺語文化，北科大錄製了臺文語音語料庫(Taiwanese Across Taiwan corpus, TAT corpus) (Liao *et al.*, 2020)並舉辦了臺語語音辨識競賽(Formosa Speech Recognition Challenge 2020, FSR-2020)，比賽項目則分為三大類型，分別為將臺語語音資料辨識為繁體中文字的 Track1（即翻譯任務），以及將其辨識為臺文的 Track2 與臺灣閩南語羅馬字拼音的 Track3。在這個研究中，我們以將臺語辨識為臺文的 Track2 與臺灣閩南語羅馬字拼音的 Track3（臺羅數字調，不考慮音調部分）為研究主題，嘗試使用目前主流之端對端(End-to-end)語音辨識器進行臺語語音辨識之任務。

2. 方法 (Methods)

2.1 競賽2: 臺文辨識 (Track2: Taiwanese Southern Min Recommended Characters Recognition)

由於近年在語音辨識上基於深度模型之自動語音辨識器蓬勃發展 (Watanabe *et al.*, 2017; Gulati *et al.*, 2020; Dong *et al.*, 2018)，在許多資料集上都達到比傳統模型更好的成果，也因此這些方法逐漸取代了傳統的機率式模型（例如隱藏式馬可夫(Hidden Markov Model, HMM)模型(Bahl *et al.*, 1986)）。因此，在本研究中，我們採用基於連結時序分類(Connectionist Temporal Classification, CTC)以及注意力機制(Attention Mechanism)的混合架構(Watanabe *et al.*, 2017; Watanabe *et al.*, 2018)，探討當前語音辨識模型於臺文語音辨識的成效。

2.1.1 模型架構 (Model Architectures)

我們採用目前端對端語音辨識模型的主流架構，即基於 CTC 以及 Attention 的混合模型 (Hybrid CTC-Attention based Models) (Watanabe *et al*., 2017)，並使用端到端的語音處理工具 ESPnet(End-to-end Speech Processing Toolkit) (Watanabe *et al*., 201 8)進行實作。ESPnet 可用來完成如語音辨識、語音合成等任務，在模型方面 ESPnet 提供了很多經典的模型供開發者做使用，如：Transformer (Watanabe *et al*., 2017; Vaswani *et al*., 2017; Dong *et al*., 2018)、Conformer (Gulati *et al*., 2020)、RNN (Elman, 1990; Hochreiter & Schmidhuber, 1997; Cho *et al*., 2014)等架構；在模型後端的神經網路框架有 Pytorch 及 Chainer 兩個版本可供選擇；在語音特徵預處理方面，使用 Kaldi 來提取語音特徵；做為一個主流的開源軟體，ESPnet 也提供了許多常見語音辨識資料集之範例腳本（如：AISHELL、Librispeech、WSJ 等）。因此基於 ESPnet，可以藉由修改各個部件來完成一套新的模型架構，也可以修改範例腳本，就可以基於新的資料集訓練出一套新的語音辨識器。

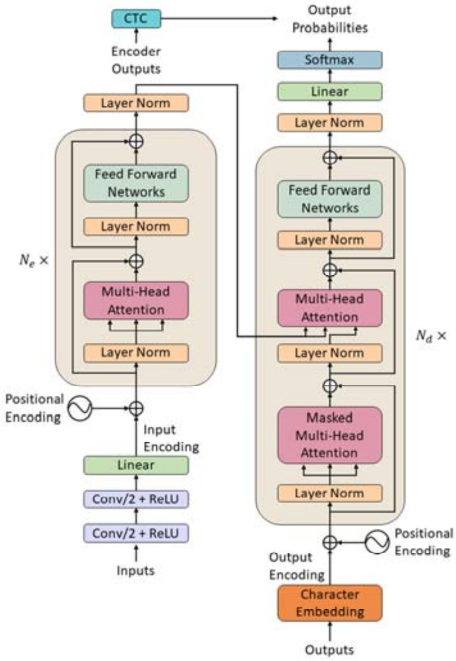

圖 1. ESPnet 中的 Transformer 模型架構圖。
[Figure 1. Model architecture of the ESPnet.]

圖 1 為我們採用的 ESPnet 架構圖，圖中左方為編碼器(Encoder)右方為解碼器(Decoder)。由架構圖可觀察出與 Speech Transformer (Dong *et al*., 2018)型相比，此模型在編碼器的輸出部分有加上 CTC 的模組。在編碼器部分，ESPnet 提供了各式各樣的神經網路模組供選擇，如：全連接神經網路或是卷積神經網路等，預設情況下 ESPnet 會使用兩層步幅為 2 的卷積層對輸入語音訊號進行降維並接著數層 Transformer 抽取出高層次的表示法，最後透過一個全連接層並計算輸出與訓練資料標籤的連結時序分類誤差，此外

依照原始 Transformer 解碼器的設計，會將編碼器的輸出傳遞至解碼器，作為解碼器中第二個多頭注意力(Multi-Head Attention)層中的鍵值(Key)及值(Value)。在解碼器輸入部分，ESPnet 提供了一般的嵌入層(Embedding)及全連接層供開發者選擇，此部分主要是用來將訓練資料標籤內的每個字元或詞轉變為向量的形式，以便進行後續的運算，最後 ESPnet 會將訓練資料標籤先進行標籤平滑化(Label Smoothing)，再接著計算解碼器輸出與訓練資料標籤的相對熵(KL Divergence)誤差，以利模型在後續進行參數更新時使用。

更明確地，ESPnet 在訓練時，會將 CTC 誤差 $\mathcal{L}^{ctc}$ 與解碼器的相對熵誤差 $\mathcal{L}^{att}$ 透過超參數 α 進行線性插值：

$$\mathcal{L} = \alpha\mathcal{L}^{ctc} + (1-\alpha)\mathcal{L}^{att} \tag{1}$$

而至於兩者之間的權重 α 為開發者可自行調整的參數，後續會利用相加過後的誤差來計算梯度並更新模型的參數；預測階段 ESPnet 使用 Beam Search 的方式進行解碼，並且在解碼階段可選擇是否要套用 RNN 語言模型來輔助模型進行預測。

2.1.2 ROVER

語音辨識模型可以透過許多方法來優化辨識的結果，例如透過提取更好的特徵或降低雜訊影響力的方式來優化模型的表現，亦或是混合多種模型的輸出結果，透過投票和重新評分的流程來決定最後的模型輸出。因此，在本研究中，我們首先訓練了許多不同架構與參數的模型後，再利用 ROVER(Recognizer Output Voting Error Reduction) (Fiscus, 1997) 將模型間的輸出結果進行整合。ROVER 能夠透過投票和重新評分的過程來決定最後的輸出，進而降低預測結果的錯誤率。圖 2 為 ROVER 的示意圖。

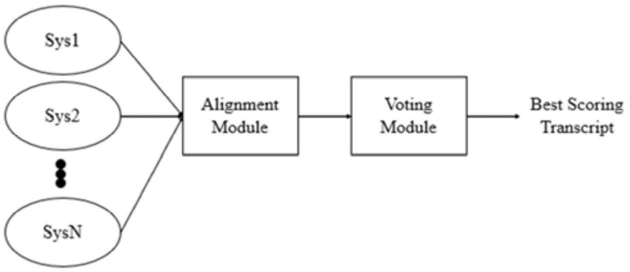

圖 2. ROVER 系統示意圖。
[Figure 2. System diagram of the Rover.]

ROVER 由兩大模組所組成，分別為對齊模組(Alignment Module)以及投票模組(Voting Module)。對齊模組會將各個辨識系統產生的結果進行對齊，接著投票模組會對每一個候選字計算一個介於 0 至 1 之間的信心分數(Confidence Score)，信心分數為 0 表示沒有任何信心，1 則表示有完全的信心。最後，投票模組有三種投票評分方式，分別為完全依靠詞頻(Word Frequency)進行投票、同時考慮詞頻與平均信心分數以及考慮詞頻與最高信心分數等方法。在本研究中，我們僅使用詞頻作為投票的依據。

2.2 競賽3: 臺羅拼音辨識 (Track3: Taiwan Minnanyu Luomazi Pinyin Recognition)

在臺羅拼音方面,本研究則嘗試使用 Kaldi 語音辨識工具中的 HMM-TDNN 混合模型，以 Tedlium 的範例為基礎修改,並使用了四連語言模型(4-gram)與 RNN 語言模型進行兩種實驗。模型架構圖如圖 3 所示。

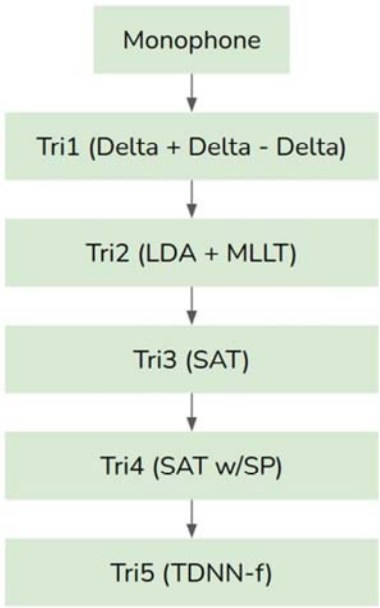

圖 3. HMM-TDNN 模型架構圖。
[Figure 3. Model architecture of the HMM-TDNN.]

3. 實驗 (Experiments)

3.1 臺文辨識 (Taiwanese Southern Min Recommended Characters Recognition)

在 Track2 的臺文辨識任務上，我們使用 ESPnet 所提供之基於 CTC 以及 Attention 的混合模型進行臺文辨識，輸入語音使用 80 維的 Fbank，模型編碼器由 12 層 Transformer 編碼器所組成（N_e=12），模型解碼器則由 6 層 Transformer 解碼器（N_d=6），多頭注意力機制數量採用 4 個 head,Transformer 中的 FFN 維度為 2048,誤差函數的超參數權重 α 為 0.3，Dropout 比率設定為 0.1，學習優化器採適應性矩估計演算法(Adam)對模型進行參數更新，並訓練 50 個 Epoch，使用 Beam Search 演算法預測時，Beam Size 設為 10，語言模型權重預設為 0.7，實驗以上述設定為基礎並嘗試各種不同的架構與設定來改善錯誤率。

在 FSR-2020 競賽期間我們將一開始收到的訓練資料(TAT-Vol1)隨機各取 10%的資料為初步實驗之開發集與測試集，進行最終實驗時我們將比賽中期釋出之 Pilot-test 測試語料加入訓練資料集中，公視臺語台訓練語料則隨機切出原始開發集與測試集之 1.5 倍左右數量之資料，加入原本的開發集與測試集中，其餘資料則做為訓練資料使用。

初步實驗使用尚未加入 Pilot-test 測試語料與公視臺語台訓練語料的資料集進行實驗，語料的統計資料如表 1 所示。

表 1. 資料集統計資訊。
[Table 1. Statistics of the dataset]

		資料數量	備註
初步實驗	訓練集	13,933	
	開發集	1,740	
	測試集	1,728	
中期實驗	訓練集	134,678	初步實驗訓練集加上公視臺語台 120,745
	開發集	1,740	
	測試集	1,728	
最終實驗	訓練集	132,142	初步實驗訓練集加入 Pilot-Test 測試語料 2,664 與公視臺語台 115,545
	開發集	4,340	初步實驗開發集加入公視臺語台 2,600
	測試集	4,328	初步實驗測試集加入公視臺語台 2,600

初步實驗集的結果如表 2 所示，使用 Transformer 架構的模型可在測試集上達到 21.7%的字元錯誤率(CER)，在使用預訓練(Pre-train)於 AISHELL-1 與 Librispeech 資料集（以字元為單位）的模型參數進行微調(Fine-tune)後，在測試集上的錯誤率可下降至 16.5%，若加上速度擾動(Speed Perturbation)可進一步下降到 15.3%，此外若使用字節對編碼(Byte Pair Encoding, BPE)為單位進行預訓練，錯誤率可再下降至 14.9%，此模型亦為繳交 Pilot-test 之模型，在 Pilot-test 所測得之錯誤率為 25.22%。

表 2. 初步實驗資料集之字元錯誤率(CER)。
[Table 2. Character error rate (CER) on the preliminary dataset.]

模型	開發集	測試集
Transformer Model	21.8	21.7
Pre-train Model (char)	16.0	16.5
+ Speed Perturbation	15.0	15.3
Pre-train Model (BPE)	**14.7**	**14.9**

於比賽中期我們嘗試將公視臺語台訓練語料先全部加入訓練集進行訓練(即中期實驗資料集)，結果如表 3 所示，由於在我們的初步實驗中，使用速度擾動與使用字節對編

碼預訓練均能有效改善錯誤率，所以這組實驗均預設地加入這兩種方式進行訓練。

表3. 中期實驗資料集之字元錯誤率(CER)。
[Table 3. Character error rate (CER) on the intermediate dataset.]

模型	開發集	測試集
Transformer Model	16.0	16.3
Pre-train Model (BPE)	**14.8**	**15.1**

有鑑於公視臺語台訓練語料可能有較多雜音或背景音與一開始釋出的訓練語料差異較大，因此最終實驗中也將公視臺語台訓練語料加入至最終實驗的開發集與測試集之中（即最終實驗資料集）。在表 4 實驗結果中，我們嘗試使用不同架構模型（未進行預訓練且未使用語度擾動），可發現 Transformer 架構的表現比傳統 RNN 架構優秀很多，而後續推出之 Conformer 架構也能獲得比 Transformer 架構更好的結果，不過 Conformer 架構的缺點則為訓練時間會比 Transformer 架構來的更久。

表4. 最終實驗資料集之字元錯誤率(CER)。
[Table 4. Character error rate (CER) on the final dataset.]

模型	開發集	測試集
Transformer	26.4	26.4
Conformer	**25.8**	**25.6**
RNN	29.2	29.5

接著，我們以 Transformer 模型為主，再加上以字節對編碼為單位進行預訓練於 AISHELL-1 與 Librispeech 資料集，並且在預測時調整不同語言模型權重之結果於表 5 所示。實驗結果顯示，當語言模型權重為 0.2 時可得到字元錯誤率 20.8%為最佳成果。我們也比較了 Conformer 架構與 Transformer 架構在未進行預訓練並在未使用語速擾動的辨識成效，實驗結果如表 6 所示。我們發現當權重設定為 0.1 時，可以獲得最好的結果，並且語言模型權重的影響與表 5 之結果非常接近。

表5. Transformer 模型使用不同語言模型權重之字元錯誤率(CER)。
[Table 5. Character error rate (CER) for Transformer model with different language model weights.]

語言模型權重	開發集	測試集	初步測試集
0	20.9	21.2	9.9
0.1	**20.6**	20.9	9.6
0.2	20.7	**20.8**	**9.5**
0.3	20.8	20.9	9.6
0.5	21.6	21.7	9.9
0.7	24.5	24.3	10.6

表 6. 不同模型架構使用不同語言模型權重之字元錯誤率(CER)。
[Table 6. Character error rate (CER) for Transformer model and Conformer model with different language model weights.]

Transformer Model		
語言模型權重	開發集	測試集
0.1	**22.4**	**22.4**
0.4	24.1	23.9
0.7	26.4	26.4
Conformer Model		
語言模型權重	開發集	測試集
0.1	**23.4**	**23.2**
0.7	25.8	25.6

由於實驗訓練了多個模型，所以我們嘗試將其中三個最佳的模型以 ROVER 結合，試著得到更好的結果，使用模型包括：同時使用預訓練與語速擾動之 Transformer 模型、僅使用語速擾動之 Transformer 模型、未使用語速擾動之 Conformer 模型，經過 ROVER 產生之結果在測試集上字元錯誤率可下降至 20.3%。於 Final-test results 我們選擇繳交表五實驗中最佳的模型（即語言模型權重為 0.2）作為 X21，以及使用 ROVER 結合模型預測的結果作為 X22，序號 X21 與 X22 的系統在最後 Final-test 中分別獲得 67.7%與 66.1%的字元錯誤率。

3.2 臺羅拼音辨識 (Taiwan Minnanyu Luomazi Pinyin Recognition)

在臺羅拼音辨識的任務上，我們將比賽提供的資料切割成訓練集、開發集與測試集，其統計資料如表 7 所示；語音辨識模型採用 Kaldi 之 HMM-TDNN 混合模型，並使用四連語言模型(4-gram)與 RNN 語言模型，兩種實驗結果如表 8 所示。實驗結果中， Baseline 模型為使用主辦單位提供之基礎系統進行訓練。在 Final-test 中，我們系統的字元錯誤率為 28.6%，與訓練時的成果有一段差距。可能的原因是因為 Track3 中無法使用公視臺語台、民視電視劇資料集，只使用了 TAT-Vol1 進行訓練，而比賽最終測試集包含合成噪音的語音，實驗中只有乾淨的音檔，並且沒有對噪音進行額外的處理。因此在資料量少，且沒有複雜的前處理下，我們在 Final-test 中並沒有獲得特別突出的成果。

表 7. 資料集統計資訊。
[Table 7. Statistics of the dataset.]

	資料數量
訓練集	18,101
開發集	2,282
測試集	2,218

表 8. 臺羅拼音辨識之字元錯誤率(CER)。
[Table 8. Character error rate (CER) for Track3.]

模型	開發集	測試集
Baseline	4.97	6.13
N-gram LM	2.77	3.75
RNN LM	2.85	3.91

4. 結論 (Conclusions)

在臺文辨識的 Final-test 上，我們所得到的錯誤率為 66.1%(X22)、67.7%(X21)，與測試在自行分割的測試集上相差許多；臺羅拼音辨識方面，我們所得到的錯誤率為 28.6%。後續觀察 Final-test 的音檔可發現，也許是因為 Final-test 中的音檔與訓練資料集形式差異較大所導致，先前所釋出音檔多為特地錄製的單人語料與公視電視台語料，而 Final-test 音檔不但有廣播、多人對話，有些語料還有噪音或是聲音忽大忽小的問題，且每段的說話時間也較長，由於未考慮這些特殊狀況來進行預處理，模型在預測上容易無法成功辨識的情形，所以導致結果不是很好。

參考文獻 (References)

Bahl, L., Brown, P., De Souza, P., & Mercer, R. (1986). Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'86.)*, *11*, 49-52. https://doi.org/10.1109/ICASSP.1986.1169179

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., …Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In arXiv preprint arXiv:1406.1078

Dong, L., Xu, S., & Xu, B. (2018).Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, 5884-5888. https://doi.org/10.1109/ICASSP.2018.8462506

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179-211. https://doi.org/10.1016/0364-0213(90)90002-E

Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proceedings of 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, 347-354. https://doi.org/10.1109/ASRU.1997.659110

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., …Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. In arXiv preprint arXiv:2005.08100

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Liao, Y.-F., Chang, C.-Y., Tiun, H.-K.., Su, H.-L., Khoo, H.-L., Tsay, J. S., Tan, L.-K., Kang, P., Thiann, T.-g., Iunn, U.-G., Yang, J.-H.,…Liang, C.-N. (2020). Formosa Speech Recognition Challenge 2020 and Taiwanese Across Taiwan Corpus. In *Proceedings of 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA '20)*, 65-70. https://doi.org/10.1109/O-COCOSDA50338.2020.9295019

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., …Polosukhin, I. (2017). Attention is all you need. In arXiv preprint arXiv:1706.03762

Watanabe, S., Hori, T., Kim, S., Hershey, J. R., & Hayashi, T. (2017). Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, *11*(8), 1240-1253. https://doi.org/10.1109/JSTSP.2017.2763455

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., …Ochiai, T. (2018). Espnet: End-to-end speech processing toolkit. In arXiv preprint arXiv:1804.00015

Computational Linguistics and Chinese Language Processing
Vol. 26, No. 1, June 2021, pp. 43-78

Textual Relations with Conjunctive Adverbials in English Writing by Chinese Speakers: A corpus-based Approach

Tung-Yu Kao[*] and Li-Mei Chen[*]

Abstract

The study aims to investigate the use of conjunctive adverbials (CA, hereafter) performing various textual relations in the English writing by Chinese speakers across genres and over time. To begin with, a corpus of one million word was compiled and the corpus interface was constructed. Later, 45 pieces of writing by 5 college students during 4 semesters were selected for data annotation and analysis, with each student contributing 9 pieces for 9 text genres. The results show that there exists a distribution norm of CA-performed textual relations based on CA occurrence frequency and that the distribution is independent of genre and time effects. Compared with literature, the found distribution is also considered free from the first language influence. This suggests that the found distribution is a mental representation of mature human cognition, underlying English writing on global and coherent levels. Therefore, the found distribution is of great potential for developing automatic tools of discourse diagnosis.

Keywords: Conjunctive Adverbial, Textual Relation, Text Genre, English Writing, Corpus Compilation, Automatic Discourse Diagnosis.

1. Introduction

Implemented in August 2019, Curriculum Guidelines of 12-Year Basic Education (MOE, 2014) has ushered in a new era of the English education in Taiwan. In the past, the English education in school was under severe criticism for overemphasizing sentential grammaticality and understating discoursal and pragmatic aspects. The concern has received attention and been addressed. In the spirits of the curriculum reform, the Guidelines for the English education (MOE, 2018) specifies that students should be taught to identify text genres through

[*] Department of Foreign Languages and Literature, National Cheng Kung University
E-mail: dodofishletter@hotmail.com; leemay@mail.ncku.edu.tw
The author for correspondence is Li-Mei Chen.

text features, pay attention to cohesion and coherence across sentences, and employ reading strategies, such as skimming and inferring, to comprehend the text as a whole. In other words, the focus of the English education in Taiwan evolves to direct students to the importance of textual relations on the discourse level.

In fact, in higher education, the emphasis of textual relations in classes for English writing had been well addressed. It is found that most writing textbooks (Connelly, 2013; Langan, 2010; Lannon, 2007; Morenberg & Sommer, 2008; Reid, 2000; Smalley *et al*., 2012; Wyrick, 2008) introduce various genres by indicating that some conjunctive adverbials (CAs, hereafter) are more prominent in certain genres than other CAs based on the textual relations they perform. For instance, CAs, such as *firstly*, *next*, and *in addition*, are thought to appear more in the process genre to show progressive relations across sentences or paraphrases. All things considered, learning to recognize and employ CAs that signal textual relations in genres can be said to become a highlighted component throughout the English education in Taiwan.

Also, to respond to the enormous demand of English writing by native speakers (NS, hereafter) and by non-native speakers (NNS, hereafter), writing-assisting online platforms and software packages, such as *My Access!*, *Grammarly*, *StyleWriter*, and *WhiteSmoke*, have been developed for the purpose of automatically facilitating and advising writers to compose better. As seen in Table 1, *My Access!* is a process-oriented writing platform where people are guided through scaffolds, like brainstorming charts and revision requests, for contents development, and provided with human-graded scores and feedback as reference. The other three are product-checking writing software to proofread writing forms, including punctuation, spelling, and sentential grammaticality, for correctness, and to offer possible alternatives for more concise sentence rephrasing and better word choice.

Table 1. Functions offered by the four writing-assisting software.

Functions / Tools	Form Checker	Style Options	Writing Scaffold	Score & Feedback	Unique Feature
My Access! [1]	+	–	+	+	Previous writing pieces analyzed based on grammar mistakes
Grammarly[2]	+	+	–	–	Different options provided based on desired tones
StyleWriter[3]	+	+	–	–	Statistics, such as word count and sentence length, provided
WhiteSmoke[4]	+	+	–	–	Full-text and word-to-word translation available

Note: [1]https://www.myaccess.com/myaccess/do/log
[2]https://www.grammarly.com/
[3]https://www.editorsoftware.com/stylewriter.html
[4]https://www.whitesmoke.com/

With these textual foci in the English education in Taiwan and technology advances in the English writing, however, few empirical studies has been conducted after Chen's (2006) seminal work to examine whether the difference in the occurrence of CAs across different genres is on a significant level, and existing writing tools are still inadequate to automatically diagnose a writing text and generate comments on the discourse level. In a preliminary database search on Linguistics and Language Behavior Abstracts (LLBA), the results for peer-reviewed articles after 2007 with *conjunctive adverbial* and *writing* in the abstract produced 92 entries. On closer inspection, none investigated various textual relations across different genres. Although there were studies explored the use of CAs, the CA items were pre-selected and belonged to the same textual relation, such as Phoocharoensil (2017) exploring resultive CAs *THUS*, *THEREFORE*, *HENCE*, and *SO* in written academic English. In another database search on Academic Search Complete (EBSCOhost), the results for peer-reviewed articles after 2007 with *automatic, evaluation,* and *discourse* in the abstract produced 88 entries. On closer inspection, the only study dedicated to computational text-level discourse analysis is Morey, Muller, and Asher' (2018) study. Yet, their study was based on Rhetorical Structure Theory (RST), rather than textual relations performed by CAs.

Therefore, the present study aims to take a corpus approach to investigate (1) whether the distribution of CA-performed textual relations in the English writing by Chinese speakers is significantly subject to text genres, and (2) whether it significantly varies over time with learning and practicing. Ultimately, the data collected and the results are hoped to serve as training data and calculating principles for developing automated discourse-evaluating application.

2. Literature Review

This section consists of two parts, with the first part focusing on the working definition of CAs, short for conjunctive adverbials, employed in the present study and the other reviewing previous studies on their distribution in the English writing by native and non-native speakers.

2.1 Working Definition of CAs (Conjunctive Adverbials)

In Halliday and Hasan's (1976) cohesion framework, CAs are one type of cohesion to achieve textual coherence by which sentences are grouped together and considered an integrated discourse unit. The type of cohesion differs from other types in the way it functions to make connections among sentences. CAs relate sentences by providing one possible interpretation to confine the effect of sentences on one another, rather than using anaphoric relation to ensure the involvement of the same topic in sentences.

For example, the two sentences in (1) are regarded as one unit for the pronoun in the second sentence refers back to the subject in the first sentence and establishes a link between the two sentences. Unlike both sentences in (1) staying with the topic of the person, sentences in (2) deal with different topics but are still viewed as a unit because the CA, *however*, offers one kind of textual relation to denote how the two propositions in (2) are related to each other. From the discussion, CA, therefore, can be defined as one text-creating mechanism indicating the inter-sentential textual relation, and accordingly exclude the discussion of coordinators or subordinators, which signal the intra-sentential textual relation.

(1) *Barack Obama* was inaugurated as the President of the United States on January 20, 2009. *He* is the first African American president in the history of the country.

(2) May is the plum rains season in Taiwan. *However*, the rainfall this year reaches a historic minimum.

Nevertheless, the definition is not exclusive enough. As seen in (3a) and (4a), both *however* and *later* indicate how the second sentence is related to the first sentence in respective examples, with the former yielding a contrastive effect and the latter designating the temporal order. Yet, not both are considered CAs. According to Quirk, Greenbaum, Leech, and Svartvik (1985), one common feature shared by CAs, sentence adverbials in Quirk *et al*'s term, is that the type of cohesion cannot occupy the focus of a cleft sentence. In this sense, after tested in (3b) and (4b), *however* remains a CA while *later* would be excluded from the scope of the present study.

(3) a. When the Tae Kwon Do contestant Li-wen Su sprained her knee in Olympics, people thought she would quit the contest. *However*, she continued fighting to the end.

b. *It is however that she continued fighting to the end.

(4) a. Landing on the moon was first ridiculed as an impossible mission. *Later*, people realized this could really work.

b. It is later that people realized this could really work.

Given the semantic and syntactic criteria for defining CAs, an additional criterion is employed in the present study. According to Halliday and Hasan (1976), CAs fall into three kinds of language form: adverbs, prepositional phrases, and prepositional expressions with reference items, as presented in (5a) to (5c), respectively. It is clear that (5a) and (5b) are linked to the first sentence because of the adverb in (5a) and the prepositional phrase in (5b) specifying textual relations between sentences. However, the link between (5c) and the first sentence is established based mainly on the presence of the reference in the prepositional expression. The language form that Halliday and Hasan also regard as CAs works more like lexical cohesions than CAs. Therefore, the third criterion supplemented in the present study is that a CA must be lexicalized and self-contained. In other words, the present study only investigates CAs in the form of adverbs and prepositional phrases.

(5) The captain had steered a course close in to the shore.

a. *Therefore*, they avoided the worst of the storm.

b. *As a result*, they avoided the worst of the storm.

c. *As a result of this*, they avoided the worst of the storm.

Ultimately, the working definition of CAs in the present study is as follows, and Table 2 shows how the three criteria delimit the investigating scope in the present study.

Criterion 1: A conjunctive adverbial must semantically indicate the relation between the sentences before and after it.

Criterion 2: A conjunctive adverbial must be syntactically forbidden to be the focus in a cleft sentence.

Criterion 3: A conjunctive adverbial must be lexicalized and self-contained.

Table 2. Working definition delimiting the investigated CAs.

The CA forms in Halliday and Hasan (1976)	Examples	Tested based on the three criteria	To be examined in the present study
Adverbs	therefore	Satisfying the criteria	Yes
Prepositional phrases	as a result	Satisfying the criteria	Yes
Prepositional expressions with reference items	as a result of that	Violating criterion 3	No

2.2 Previous Studies on Textual Relations Performed by CAs

The significance of CAs lies in the fact that they direct the interpretation among sentences in text, which leads to the attempt to classify textual relations explicitly indicated by conjunctive adverbials. According to Halliday and Hasan (1976), *Additive*, *Adversative*, *Causal*, and *Temporal* were the four types of textual relations regulated, with various subdivisions in each type. Later, the various subdivisions were collapsed, and the taxonomy was simplified by Celce-Murcia & Larsen-Freeman (1999).

Table 3. CA taxonomy in the literature.

Systems	Researchers	Types
Four-type classifying system	Halliday & Hasan (1976)	Additive, Adversative, Causal, Temporal
	Celce-Murcia & Larsen-Freeman (1999)	Additive, Adversative, Causal, Temporal
More-type classifying system	Quirk *et al*. (1985)	Listing, Summative, Appositional, Resultive, Inferential, Contrastive, Transitional
	Biber *et al*. (1999)	Enumeration, Addition, Summative, Appositional, Result/Inference, Contrast/Concession, Transitional

Compared with Celce-Murcia and Larsen-Freeman collapsing subdivisions, Quirk *et al*. (1985) revised Halliday and Hasan's four-type system as a system of seven types, namely, *Listing*, *Summative*, *Appositional*, *Resultive*, *Inferential*, *Contrastive*, and *Transitional*. Similarly, Biber *et al*. (1999) developed their own classifying version, which was very much the same with Quirk *et al*.'s except separating *Listing* in Quirk *et al.* as *Enumeration* and *Addition*, changing *Contrastive* as *Contrast/Concession*, and combing *Resultive* and *Inferential* into *Result/Inference*. Table 3 summarizes CA taxonomy based on the four-type and the more-type classifying systems.

With the four-type and the more-type classifying systems, empirical studies involving investigation into the distribution of CA-performed textual relations in the English writing are reviewed. Table 4 and 5 present studies that employed the four-type framework of CAs for analysis, with the former based on the English writing by NS and the latter based on the English writing by NNS. As shown in Table 4, the distribution patterns of CA-performed textual relations in Field and Yip (1992) and in Chen (2006) are identical. CAs of *Adversative* occur the most frequently, followed by *Additive*, *Causal*, and *Temporal* in a descending order. This may suggest that there exists a distribution norm in NS cognition, and that the norm is independent of genre and time influences.

In contrast, the writing by NNS does not present such a distribution norm of textual relations carried out by CAs. As seen in Table 5, the five studies present four distribution patterns; in other words, the results in these studies differ from one another. In addition, genre and time influences fail to account for the lack of distribution pattern consistency. Consider genre influence. Field and Yip (1992) as well as Liu and Braine (2005) investigated the same writing genre, and so did Xie (2014) and Huang (2018). Yet, neither sets of studies found the same distribution pattern. Now, consider time influence. Field and Yip (1992) as well as Xie (2014) collected data from the same time period, and so did Liu and Braine (2005) as well as Huang (2018). Again, neither sets of studies found the same distribution pattern.

Table 4. Distribution of textual relations in the NS writing based on four-type framework.

Researchers	Field & Yip (1992)	Chen (2006)
Framework	H & H (1976)[1]	C & L (1999)[2]
Genre	Argumentation	Research articles
Data Source	High school students in Sydney	Journal articles on TESOL
Distribution of Textual Relations	Adversative >[3] Additive > Causal > Temporal	Adversative > Additive > Causal > Temporal

Note: [1]Halliday & Hasan (1976)
[2]Celce-Murcia & Larsen-Freeman (1999), the simplified framework of Halliday and Hasan
[3]The symbol “>” means “more occurring frequencies than.”

To account for the lack of distribution pattern consistency, first language background is suggested to be the cause. After comparison, it is found that the distribution pattern exhibited in the NNS writing in Field and Yip (1992) is the same as that in the NS writing shown both in Field and Yip (1992) and in Chen (2006). The consistency may be explained by the fact that the so-called NNS group in Field and Yip (1992) should be considered Chinese-English bilingual natives. They lived in once-UK-colonized Hong Kong and were immersed in the English-speaking environment growing up. Consequently, they may share the same distribution pattern with NS in cognition in terms of English writing. The language background also account for why the NNS writing in Liu and Braine (2005) exhibits the same distribution pattern as that in Chen (2006) since Mandarin is the first language for both data sources.

Table 5. Distribution of textual relations in the NNS writing based on four-type framework.

Researchers	Field & Yip (1992)	Liu & Braine (2005)	Chen (2006)	Xie (2014)	Huang (2018)
Framework	H & H (1976)	H & H (1976)	C & L (1999)	H & H (1976)	H & H (1976)
Genre	Argumentation	Argumentation	Various kinds	Exposition	Exposition
Data Source	High school students in Hong Kong	College students in Beijing	MA TESOL students in Taiwan	High school students in Taiwan	College students in Taiwan
Distribution of Textual Relations	Adversative> Additive> Causal> Temporal	Additive> Causal> Temporal> Adversative	Additive> Causal> Temporal> Adversative	Additive> Temporal> Causal> Adversative	Temporal> Additive> Adversative> Causal

However, for the four studies with data sources taking Chinese as the mother tongue, the distribution patterns in Xie (2014) and in Huang (2018) still differ from the pattern exhibited in Liu and Braine (2005) and in Chen (2006). The difference may lie in writing length. The length of one piece of writing collected in Xie (2014) is 120 to 170 words, and that in Huang (2018) is 150 to 200 words. Since the writing collected in Chen (2006) is research-related, including literature reviews, research proposals, and pedagogical “how-to” papers, the average length is much longer, between 3000 and 4000 words. Given the above observation and discussion, it might be inferred that while time and genre differences do not impact the distribution pattern of textual relations, first language and writing length might have a role in affecting it.

Table 6. Distribution of textual relations based on more-type framework.

Researchers	Tankó (2004)		Altenberg & Tapper (1998)		Shen (2006)	
Framework	Quirk *et al*. (1985)		Quirk *et al*. (1985)		Biber *et al*. (1999)	
Genre	Argumentation		Argumentation		Research articles	
Data Source	NS: No data	NNS: Hungarian college students	NS: College students	NNS: Swedish college students	NS: Journal articles on TESOL	NNS: Conference papers by Taiwanese
Distribution of Textual Relations		Listing> Resultive> Contrastive> Summative> Appositive> Inferential> Transitional> Corroborative (0 case)	Contrastive> Resultive> Listing> Appositive> *Corroborative*> Summative> Transitional (0) Inferential (0)	Contrastive> Resultive> Appositive> Listing> Corroborative> Summative> Transitional> Inferential (0)	Contrastive> Appositive> *Result/Inference*> Listing*> Corroborative> Transitional> Summative	Listing> Contrastive> *Result/Inference*> Appositive> Summative> Transitional> Corroborative

*The *Additive* and *Enumerative* textual relations in Shen (2006) were collapsed into *Listing* in comparison with the other studies.

Tankó (2004), Altenberg and Tapper (1998), and Shen (2006) are the studies taking the more-type system as the framework for analysis. Although the former two studies were based on Quirk *et al*.’s (1985) classification and the last was on Biber et al.’s (1999), the frameworks they employed were much the same. The only difference is the use of category names without substantial contents changes. For example, *Additive* and *Enumerative* in Shen (2006) could be collapsed and equate *Listing* in the other two studies. Moreover, the three studies all referred to Granger and Tyson’s (1996) classification and designated one more textual relation, *Corroborative*, conveying writers’ attitudes toward and comments on the text, in their frameworks. In terms of text genre and time period, Tankó (2004) as well as Altenberg and Tapper (1998) limited their data to the argumentative writing by college students with the former taking the latter's NS data as benchmark, while the data in Shen (2006) were research papers from academic journals as benchmark and conference papers by Taiwanese postgraduates as the NNS samples. The research design and results of the three studies are summarized in Table 6.

Based on Table 6, despite genre, age, and first language differences in data sources, a distribution norm of CA-performed textual relations in English writing can be identified, when textual relations are considered in groups. In the distribution norm, *Contrastive*, *Resultive*, *Listing* occur most often, *Appositive* and *Summative* ranks moderate in the order, and *Transitive* and *Inferential* are seldom used.

From the observed distribution norm, three points are induced as well. Firstly, fine classification of textual relations might be able to clarify nuances in occurrence frequency better than generic classification, and to manifest the underlying distribution norm. The inference is made due to the difference in the observed results based on the four-type and more-type frameworks. While the distribution patterns of CA-performed textual relations based on both frameworks might be not susceptible to text genre and time period, first language might have an influence on the distribution in the former framework but not in the latter. Secondly, it makes sense that Biber *et al*. (1999) combined *Resultive* and *Inferential* in Quirk *et al*.'s (1985) classification for the inferential relation has the lowest occurring ratio, e.g., the zero occurrence in Altenberg and Tapper (1998). Lastly, it is found that NS use the *Corroborative* relation in their writing more frequently than NNS do. This might originate in the fact that writers would exhibit a higher level of authority when writing in their first languages than in other languages (Chen, 2006). With corroborative adverbials serving to express writers' opinions, the kind of conjunctive device, therefore, is used more often in the NS writing for establishing authority.

3. Methodology

The present study takes a corpus-based approach to explore the distribution of CA-performed textual relations in the English writing by Chinese speakers across genres and over time. For the research goal, the section first reports how the corpus in present study was compiled, and then elaborates how the selected data were annotated with the coding scheme through the coding procedure. The section is wrapped up with an introduction to statistical measures for data analysis.

3.1 Corpus Compilation

A corpus-based approach was employed in the present study. The corpus compilation was based on the first four stages in Atkins, Clear, and Ostler's (1992) corpus building, which are specifications and design, hardware and software, data capture and mark-up, as well as corpus processing.

In the Specifications and Design stage, the corpus formation in the present study was designed based on the OLAC Metadata Set. OLAC Metadata Set is an exclusive protocol framed by the Open Language Archives Community, regulating the information for digitally archiving language resources and basing its digital storage on the XML format (Simons & Bird, 2008). It was XML format's extensible features that allowed the present study to tailor its own format for the research purpose.

Three kinds of specification were formatted in the present study. They are *Informant Background*, *Article Message*, and *Text Annotation*. *Informant Background* offers a basis for possible research directions, while *Article Message* helps select suitable materials for research analysis. Table 7 lists the complete specifications for *Informant Background* and *Article Message*.

Table 7. Specifications for Informant Background and Article Message.

Informant Background		*Article Message*	
Aspects	Specifications	Aspects	Specifications
Basic Information	Account	Author	Account
	Chinese name		Chinese name
	English name		English name
	Gender	Attribute	Academic year
	Age		Genre
Education	Vocation / Speciality		Draft
	Level of education		Title
	University		Word count
	Department	Text	Outline
Language Use	Mother tongue		English abstract
	Known languages		Chinese abstract
	Learning experience		Text body
		Revising Process	Teacher's feedback
			Author's response

Unlike the first two formations dealing with the sources of text, the third kind of formation, Text Annotation, copes with text itself and preserves the linguistic information annotating the text it is attached to. As illustrated in the following instance, the CA is tagged within a pair of pointed brackets, and the metalinguistic coding information is annotated in the first pointed brackets.

＜tag **Y Enu CA annotation=" "**＞First＜/tag＞, children who have nasal allergy always have some mental problems to some extent.

In the Hardware and Software stage, the programming language Perl was chosen to develop the corpus interface because Perl is well known for text processing, such as dealing with files, strings, and regular expressions (Suehring, 2006). The construction of the corpus interface consisted of two phases, which were requirements analysis and system implementation. The former analyzed functions the interface should offer to serve the purpose of the present study, whereas the latter used program modules to assemble the required functions.

Based on requirements analysis, researchers, students and teachers were the three identities involved. Figure 1 visualizes the layout of the interface, where functions required by different identities are specified. Note that the block highlights the functions directly related to the purpose of the present study, and only the functions in the block are further depicted in Table 8.

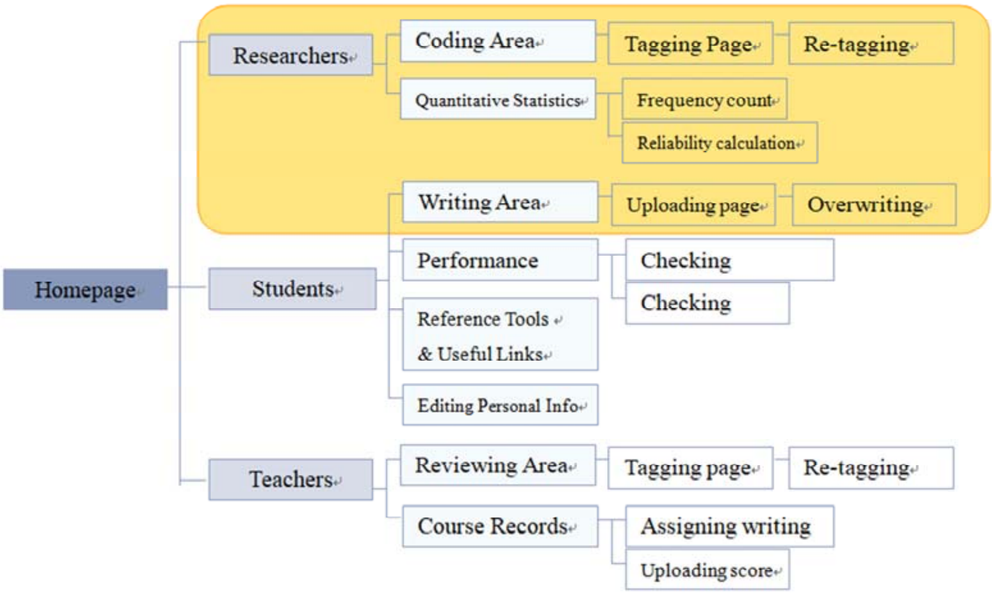

Figure 1. Interface layout.

Table 8. Description of each function.

Interface Function	Description
Tagging Page	Annotate the data with meta-information
Re-tagging	Retrieve annotated data for reviewing and revising earlier annotation
Frequency Count	Search the corpus by an analyzing code or a specific word, show all the matched cases and tally the total occurring frequency.
Reliability Calculation	Present two researchers' annotations of the same text in parallel and calculate all the combination situations of agreement and disagreement.
Uploading Page	Upload compositions.
Overwriting	Retrieve uploaded compositions for overwriting earlier drafts.

Following the requirements analysis was the system implementation of the interface. Table 9 presents and defines eleven modules that help execute functions required.

Table 9. Execution of each program module.

Program Module	Abbreviation	Execution
Highlight	High	Distinguish annotated information from raw data
Input	In	Receive the input data
Hash	Hash	Calculate agreement frequency of coders' annotation
Match	Mat	Compare annotations of coders based on units
Output	Out	Present the retrieved data on screen
Import	Imp	Import the enquired data from the corpus
Save	Sav	Save the uploaded data in the corpus
Filter	Fil	Sift data entries that match the search instruct
Login	Log	Secure the legitimacy of users
Split	Spl	Break down the text into units (sentences or words)
Tagger	Tag	Annotate the raw data with meta-information

Given the eleven program modules, Figure 2 visualizes how each function is modulized. In the figure, squares and cambers, respectively, represent the desired functions and the assembling modules. The correspondence between abbreviations and modules is presented in Table 9.

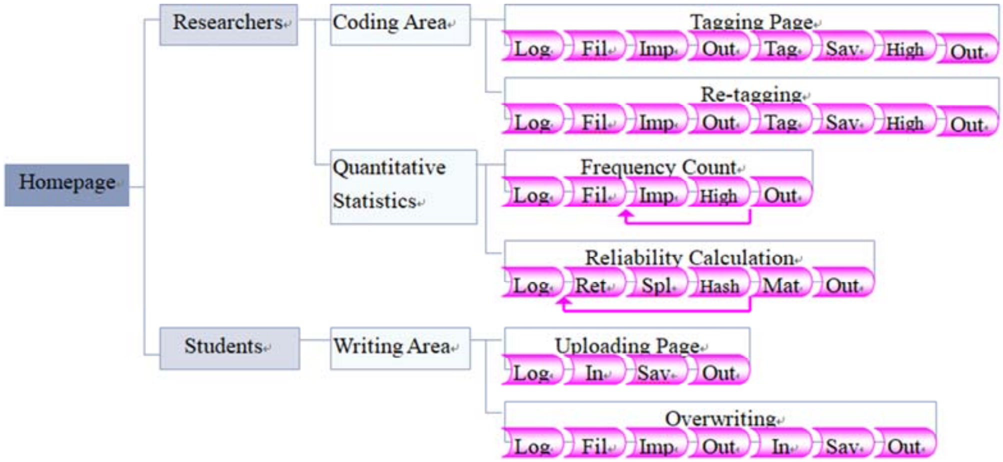

Figure 2. Assembly of modules.

To this point, the construction of the corpus interface was completed. Table 10 demonstrates how the eleven program modules assemble all the functions.

Table 10. Module Assemblage of Functions.

Function / Module	*Tagging page*	*Re-tagging*	*Frequency count*	*Reliability calculation*	*Uploading page*	*Over-writing*
Highlight	+	+	+			
Input					+	+
Match				+		
Hash				+		
Output	+	+	+	+	+	+
Import	+	+	+	+		+
Save	+	+			+	+
Filter	+	+	+			+
Login	+	+	+	+	+	+
Split				+		
Tagger	+	+				

With the corpus interface was constructed, the Data Capture and Mark-up stage ensued. Students in the Department of Foreign Languages & Literature at National Cheng Kung University in Taiwan agreed to participate in the present study, contributing their writing pieces to compile the corpus of English academic writing by Chinese NS. The genres of all the writing pieces collected belonged in 13 types, including *process*, *summary*, *essay question writing*, *cause-effect*, *comparison-contrast*, *definition*, *description*, *narration*, *classification*, *multiple strategies*, *argumentation*, *problem solving*, and *research article*. For each genre, each student wrote 3 or 4 pieces of composition, which could be independent of one another or revised drafts for the prior draft.

Figure 3 presents the *Uploading page* where students upload their writing pieces as raw data. While the system would automatically import information on *Author*, students needed to manually select labels for *Attribute*, and typed their writing in the text-editing area. Once students clicked the bottom Save and all the information would be automatically marked up and converted into machine-readable text.

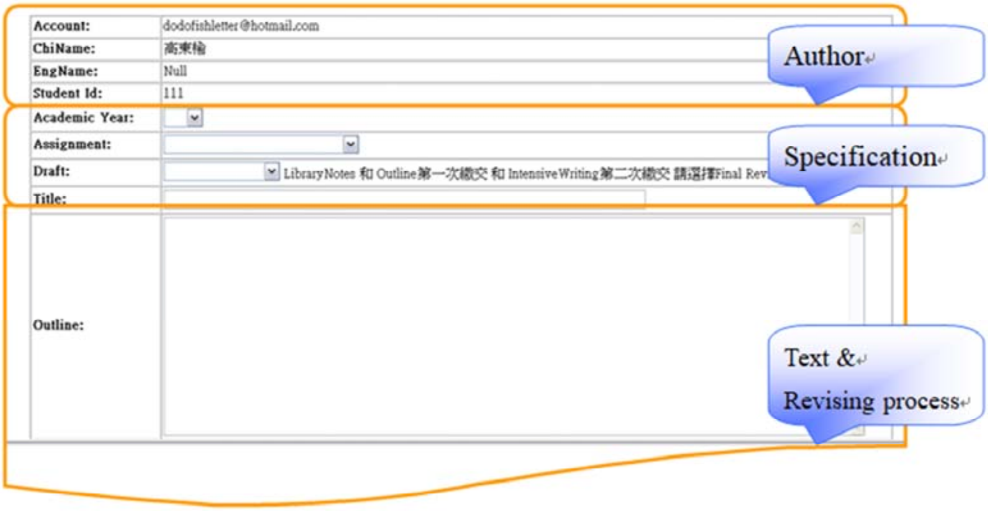

Figure 3. Uploading page.

After collecting the raw data came the Corpus Processing stage. Two most relevant data-processing functions provided on the corpus interface are *Tagging page* and *Frequency count*.

Figure 4 presents Tagging page, where researchers proceed to analyze the data. Part A is the tagger. Part B shows the coding scheme to attach to the language form in text. It is worth mentioning that the coding scheme shown on the tagger is replaceable, and can be changed according to different research purposes. Part C demonstrates the annotated text after tagging.

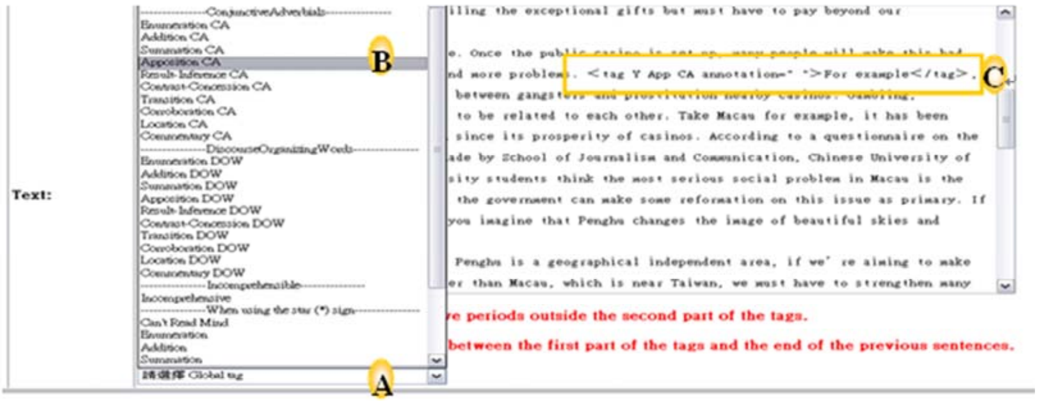

Figure 4. Tagging page.

Frequency count helps researchers obtain descriptive statistics for study results, as shown in Figure 5. Part A provides two ways to get the statistics. The frequency count can be either based on a specific tag or based on a specific word. *Subject options* aims to limit the search scope. The default search scope is the whole corpus, but researchers could narrow the search by clicking the column. Part B reports the frequency count, whereas Part C presents and highlights all the data matching the search instruction.

Figure 5. Frequency count.

3.2 Data Selection

After three years of data collection, the compiled corpus reached one million words. It consists of 2290 pieces of English compositions by Chinese speakers, belonging to 13 different genres. The total word count is 1429397 words.

Since the present study was targeted at exploring whether text genre and time period play a role in the distribution of textual relations manifested by CAs in the English writing by Chinese speakers, it was better that data provided by students covered all the genres and time periods for the purpose of minimizing individual differences and various writing instructions they received during different time periods. As a result, 45 writing pieces by 5 college students were selected to annotate and analyze. Each student contributed 9 writing pieces for 9 genres, with one piece dedicated to one genre, over 4 semesters. Table 11 shows nine genres collected, expected writing length, data word count, data sentence count, and the time period when the piece was written.

Table 11. Selected data for analysis in the present study.

Genre	Abbre-viation	Piece number	Length of a piece	Total word number	Total sentence number	Semester for data collection
Comparison-contrast	Com-Con	5	450-500	2380	114	2nd
Cause-effect	Cau-Eff	5	450-500	2494	132	2nd
Description	Des	5	450-500	2206	117	3rd
Definition	Def	5	450-500	2512	123	3rd
Narration	Nar	5	450-500	2742	177	3rd
Classification	Cla	5	700-750	3474	173	4th
Multiple-strategies	Mul-Str	5	700-750	3595	168	4th
Argumentation	Arg	5	900-1000	4682	213	5th
Problem-solving	Pro-Sol	5	900-1000	4857	232	5th

3.3 Coding Scheme and Coding Procedure

A coding scheme was developed to annotate the selected data with linguistic information. The classification of textual relations adopted in the present study followed Quirk *et al.*'s (1985) taxonomy with modification. In Quirk *et al.*'s taxonomy, there were seven types distinguished, including *Listing*, *Transitional*, *Appositive*, *Summative*, *Resultive*, *Inferential*, and *Contrastive*, whereas the present study collapsed *Resultive* and *Inferential* into one class as well as supplemented one additional semantic class, *Corroborative*. As a result, in the present study, textual relations indicated by the CAs encompasses seven types, which are *Listing*, *Transitional*, *Appositive*, *Summative*, *Resultive/Inferential*, *Contrastive*, and *Corroborative*. Table 12 lists all the textual relations with their definitions and the possible language items performing these relations.

Apart from presenting the textual relations, Table 12 also shows that one language item may serve more than one textual relation. For example, the language item *then* may perform either the *Listing* relation or *Resultive/Inferential*. In other words, the semantic coding must depend on the relation performed by the CA, not on certain fixed language items.

Table 12. Textual relations and their definitions.

Textual relation	Definition	Example
Listing	Mark the next unit of discourse with or without relative priority or temporal sequence.	first, moreover, then, in addition
Transitional	Serve to shift attention to another topic that does not follow directly from the preceding event.	meanwhile, in the meantime, now
Appositive	Provide an example or an equivalent of the preceding text.	in other words, for example
Summative	Conclude or sum up the information in the preceding discourse.	in conclusion, to summarize
Resultive/ Inferential	Mark the second part of the discourse as the result or consequence of the preceding discourse.	accordingly, then, as a result, so
Contrastive	Show incompatibility between information.	however, on the contrary, anyhow
Corroborative	Express writers' attitudes toward and comments on the text.	in fact, of course, actually

Note: The classification is based on Quirk *et al.*'s (1985) taxonomy with modification.

Another issue regarding the annotation of the textual relations is register use. Take the CA *besides* as example. While *besides* performs the *Additive* relation between sentences, it is considered spoken register and should be avoided in formal writing. Therefore, the coding scheme designed also takes register differences into account, and marks CAs as *Written* or *Spoken*. Later, to make the coding scheme applicable to the computerized interface, the scheme needs converting into the *Text Annotation* format as mentioned previously.

Table 13 presents the complete coding scheme and its electronic format. With CAs enclosed in the two pairs of pointed brackets, the word *tag* signals the beginning of a piece of annotated linguistic information while /tag the end. As for a piece of the annotated linguistic information, it constitutes three layers, separated by space. The layers specify register difference, textual relation, and supplementary annotation if necessary.

Table 13. The Complete Coding Scheme.

Textual relation		Eletronic Format	
Type	Abbreviation	Written Register	Spoken Register
Listing	Lis	＜tag W Lis annotation=" "＞ ＜/tag＞	＜tag S Lis annotation=" "＞ ＜/tag＞
Transitional	Tra	＜tag W Tra annotation=" "＞ ＜/tag＞	＜tag S Tra annotation=" "＞ ＜/tag＞
Appositive	App	＜tag W App annotation=" "＞ ＜/tag＞	＜tag S App annotation=" "＞ ＜/tag＞
Summative	Sum	＜tag W Sum annotation=" "＞ ＜/tag＞	＜tag S Sum annotation=" "＞ ＜/tag＞
Resultive/ Inferential	Res	＜tag W Res annotation=" "＞ ＜/tag＞	＜tag S Res annotation=" "＞ ＜/tag＞
Contrastive	Con	＜tag W Con annotation=" "＞ ＜/tag＞	＜tag S Con annotation=" "＞ ＜/tag＞
Corroborative	Cor	＜tag W Cor annotation=" "＞ ＜/tag＞	＜tag S Cor annotation=" "＞ ＜/tag＞

In addition to the design of the coding scheme, two pitfalls need to be tackled before data analysis as well. One is misuse of CAs, and the other is ill-formed sentences in learner writing.

Since the data sources are Chinese NS learning to write in English, misusing CAs to wrongly indicate textual relations among sentences is inevitable. When a CA misuse happens, textual coherence breaches, the reading flow is interrupted, and the text becomes difficult to comprehend. The pitfall is how to code the misused CA. The use of the CA is incorrect, so it cannot be coded with the textual relation it usually designates. To code the linguistic item with the actual textual relation between sentences is not reasonable, because the coding of the item would be researcher’s interpretation. Due to the fact that there is no way to know what textual relation the writer intended to construct between sentences, the misuse occurrence of CAs is excluded from the investigation scope.

The other pitfall is concerning ill-formed sentences in learner writing. CAs indicate textual relations across sentences, but it is found that the environment where CAs occur varies a lot. For instance, a period is used to end not only sentences but also natural constituents of language, say, a noun phrase. Sometimes it is not a natural constituent of language at all, but a grammatical mistake, such as a pseudo sentence without a finite verb. As opposed to fragmental strings of words, it is also found that sentences may not be separated properly. For example, semicolons are purposefully used to juxtapose a series of unrelated sentences which, in fact, should be severed by periods, or run-on sentences are made without proper punctuation or conjunction (Tseng & Liou, 2006). Due to this ubiquitous structural deficiency, the CAs examined in the present study are those that function and indicate textual relations across units, and a unit is decided as a group of words delimited by a period no matter whether the unit is a complete sentence, a fragment, or a multi-sentence compound.

After taking care of the pitfalls, Figure 6 visualizes the four-filter procedure of coding. Each filter identifies CAs with a linguistic label. The top filter ratifies a CA based on the working definition proposed previously. The second filter judges whether or not the CA is correctly used. The third and last filters identify its register and the textual relation it performs.

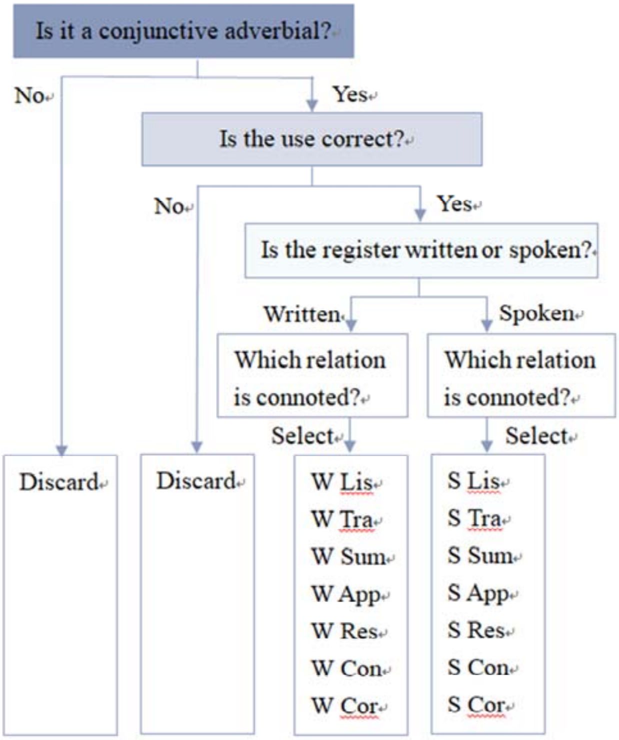

Figure 6. The coding procedure.

One researcher of the present study was the primary data annotator, responsible for annotating all the selected data for analysis. The selected data for analysis were 45 writing pieces by 5 college students. Each student contributed 9 writing pieces for 9 genres, with one piece dedicated to one genre, over 4 semesters. To ensure the reliability of data annotation, one native speaker was recruited as the inter-annotator and annotated 10% of the selected data.

The 10% of the selected data consisted of 5 pieces, with 3 pieced randomly selected from the first three semesters and 2 from the last semester. Following the coding procedure presented in Figure 6, both annotators first decided whether a lexical item satisfied the three criteria of defining a CA, proposed in Section 2.1. Then, annotators elicited their own knowledge to decide whether the CA item was used correctly and what its register was. Lastly, annotators selected a suitable tag to annotate the CA item.

The interface provides a function called Reliability Calculation, as visualized in Figure 1, to automatically report interrator agreement. The function automatically compares both annotators' annotations by examining whether both annotators tag one CA with the same textual relation. When both annotators tag one CA with the same textual relation, it is considered one match. The result shows that the matching agreement is 92.7%. Two reasons might account for the high agreement. First, the working definition of CAs provides clear semantic and syntactic criteria for CA identification. Second, since most CAs convey only one textual relation, both annotators are destined to tag most CAs with the same textual relations. Mismatch may happen only when one CA conveys more than one textual relation, yet the kind of CAs are few.

3.4 Statistical Analysis

After coding the selected data and tallying the counts, all the obtained figures were further analyzed via inferential statistical measures on SPSS to answer the two proposed research questions in the present study.

To begin with, all the raw counts of different textual relations were transformed into the-same-denominator figures to avoid the influence of writing length of the collected data in the different genres. However, while most corpus-based studies obtain the occurrence frequency ratio by using the total word count as the denominator and the CA use count as numerator, the calculation is criticized to be "fundamentally flawed" (Bolton et al., 2002) because CAs function at the discourse level. Therefore, the present study employed the unit count as denominator to normalize the occurrence frequency. As previously defined, a unit is delimited by a period regardless of the sentence structure of the unit. The unit may be a complete sentence, a cluster of words, or a multi-sentence compound.

After attaining normalized occurrence frequencies, various statistical measures were performed to answer two research questions raised. To answer research question one concerning whether genre plays a significant role in the distribution of textual relations expressed by CAs, a two-way within-subjects ANOVA was conducted, with two independent variables being textual relation and genre while the dependent variable being the CA occurrence frequency. To further examine the effect of register, the ANOVA design was calculated again, with the dependent variable becoming the written-register CA occurrence frequency. To answer research question two regarding whether time has a significant influence on the distribution of CA-performed textual relations, a two-way within-subjects ANOVA was carried out. Textual relation and genre were the two independent variables, whereas the CA occurrence frequency was the dependent variable. The ANOVA design was implemented once more to further explore the register effect. Textual relation and genre were still the two independent variables, yet the dependent variable was replaced with the written-register CA occurrence frequency. A significant level of $p<.05$ was chosen.

4. Results

This section reports whether the distribution of CA-performed textual relations varies across genres and time.

4.1 Distribution of Textual Relations Performed by CAs Across Genres

Since the use of CAs can be characterized by written and spoken registers, the distribution of CA-performed textual relations across genres is presented in three conditions, including CAs without register differentiation, written-register CAs, and spoken-register CAs.

4.1.1 Distribution of Textual Relations Through All CAs Across Genres

Table 14 presents the occurring counts of the seven CA-performed textual relations in each of the 9 genres. The raw occurring counts are signaled by *n*, and to evade the influence stemming from various writing lengths of the 9 genres, the raw occurring counts are transformed into occurring counts per 1,000 units.

Table 14. Occurring counts of textual relations performed by all CAs across genres.

Textual Relation	Com-Con		Cau-Eff		Des		Def		Nar		Cla		Mul-Str		Arg		Pro-Sol	
	n	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000
Lis	18	157.9	16	121.2	15	128.2	13	105.7	12	67.8	18	104.1	18	107.1	29	136.2	33	142.2
Tra	0	0.0	0	0.0	2	17.1	1	8.1	2	11.3	0	0.0	0	0.0	2	9.4	1	4.3
App	0	0.0	3	22.7	3	25.6	8	65.0	0	0.0	2	11.6	5	29.8	8	37.6	9	38.8
Sum	2	17.5	1	7.5	1	8.5	1	8.1	0	0.0	1	5.8	1	5.9	1	4.7	1	4.3
Res	3	26.3	7	53.0	1	8.5	4	32.5	3	16.9	4	23.1	6	35.7	12	56.3	14	60.3
Con	14	122.8	5	37.8	7	59.8	13	105.7	7	39.6	14	80.9	13	77.4	18	84.5	21	90.5
Cor	1	8.7	3	22.7	2	17.1	4	32.52	5	28.3	5	28.9	0	0.0	5	23.5	5	21.6
Unit Count	114		132		117		123		177		173		168		213		232	

Based on Table 14, a two-way within-subjects ANOVA is calculated to examine the effects of textual relation and genre. The two independent variables are textual relation and genre, while the dependent variable is the CA occurring counts per 1,000 units. The results show that there is no interaction between textual relation and genre ($F(48, 192)=1.070$, $p=0.366$) as well as no main effect from genre ($F(8, 32)=1.697$, $p=0.137$). However, there does exist a main effect from textual relation ($F(6, 24)=10.476$, $p<0.05$).

Given a main effect from textual relation, Table 15 pinpoints pairs of textual relations with significant differences, and presents related descriptive statistics. It is shown that *Listing* and *Contrastive* have the highest occurrence frequency while *Summative* and *Transitional* have the lowest. The occurrence frequencies of *Resultive/Inferential*, *Appositive* and *Corroborative* are between the two groups.

The reason of presenting the distribution with different compartments rather than in a linear sequence lies in the fact that the 3 textual-relation compartments significantly differ from one another but that there is no significant difference between textual relations within the same compartment. For example, in terms of occurring frequency, *Listing* and *Contrastive* are the highest and second highest, and both textual relations are significantly different from the other textual relations. Yet, the two are not significantly different from each other.

Table 15. Significant differences between textual relations performed by all CAs and related descriptive statistics.

Lis	Tra	App	Sum	Res	Con	Cor		Mean	SD
---	0.022*	0.035*	0.025*	0.060	0.228	0.037*	Lis	23.596	6.625
0.022*	---	0.037*	0.937	0.001*	0.005*	0.038*	Tra	1.147	0.512
0.035*	0.037*	---	0.035*	0.183	0.024*	0.526	App	4.849	1.369
0.025*	0.937	0.035*	---	0.000*	0.002*	0.002*	Sum	1.200	0.565
0.060	0.001*	0.183	0.000*	---	0.014*	0.004*	Res	6.953	0.610
0.228	0.005*	0.024*	0.002*	0.014*	---	0.003*	Con	14.153	2.199
0.037*	0.038*	0.526	0.002*	0.004*	0.003*	---	Cor	4.073	0.917

4.1.2 Distribution of Textual Relations Through Written-Register CAs Across Genres

Table 16 presents the occurring counts of the seven textual relations performed by written-register CAs in each genre. *n* designates the raw occurring counts, which are then transformed into the occurring counts per 1,000 units.

Table 16. Occurring counts of textual relations through written-register CAs across genres.

Textual Relation	Com-Con		Cau-Eff		Des		Def		Nar		Cla		Mul-Str		Arg		Pro-Sol	
	n	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000
Lis	16	140.4	11	83.3	11	94.0	10	81.3	9	50.9	18	104.1	12	71.4	24	112.7	27	116.4
Tra	0	0.0	0	0.0	0	0.0	1	8.1	0	0.0	0	0.0	0	0.0	1	4.7	1	4.3
App	0	0.0	3	22.7	3	25.6	7	56.9	0	0.0	2	11.6	5	29.8	8	37.6	9	38.8
Sum	2	17.5	1	7.6	1	8.6	1	8.1	0	0.0	0	0.0	1	5.9	1	4.7	1	4.3
Res	2	17.5	6	45.5	1	8.6	4	32.5	3	16.9	4	23.1	6	35.7	12	56.3	14	60.3
Con	13	114.0	5	37.9	6	51.3	10	81.3	5	28.3	14	80.9	13	77.4	17	79.8	20	86.2
Cor	1	8.8	3	22.7	2	17.1	4	32.5	5	28.3	2	11.6	0	0.0	5	23.5	5	21.6
Unit Count	114		132		117		123		177		173		168		213		232	

The transformed figures in Table 16 are calculated through a two-way within-subjects ANOVA to examine the distribution of textual relation performed by written-register CAs across genres. The two independent variables are textual relation and genre, while the dependent variable is the normalized occurring counts per 1,000 units. The results show that there is no interaction between textual relation and genre ($F(48, 144)=0.969$, $p=0.537$) as well as no main effect from genre ($F(8, 24)=2.062$, $p=0.082$). However, there does exist the main effect from textual relation ($F(6, 18)=8.585$, $p<0.05$).

Table 17 presents related descriptive statistics and pinpoints pairs of textual relations with significant differences. At first glance, there seems to be a distribution norm of textual relations performed by written-register CAs, much similar to that based on all CAs. Nevertheless, the seeming distribution norm is an illusion, because most textual relations do not differ from one another on a significant level. Take *Listing* and *Transitional* as an example. The former does not significantly differ from any textual relations while the latter only differs from *Contrastive* and *Resultive/Inferential*, which is very different from what happens in the distribution of textual relations performed by CAs without differentiating registers.

Table 17. Significant differences between textual relations through written-register CAs across genres and related descriptive statistics.

Lis	Tra	App	Sum	Res	Con	Cor		Mean	SD
---	0.050	0.065	0.050	0.111	0.393	0.080	Lis	20.465	6.458
0.050	---	0.084	0.759	0.008*	0.012*	0.075	Tra	0.514	0.299
0.068	0.084	---	0.056	0.283	0.045*	0.519	App	4.974	1.659
0.050	0.759	0.056	---	0.003*	0.009*	0.033*	Sum	0.391	0.231
0.111	0.008*	0.283	0.003*	---	0.024*	0.005*	Res	6.672	0.906
0.393	0.012*	0.045*	0.009*	0.024*	---	0.009*	Con	13.966	2.445
0.080	0.075	0.519	0.033*	0.005*	0.009*	---	Cor	4.028	1.169

4.1.3 Distribution of Textual Relations Through Spoken-Register CAs Across Genres

Table 18 presents the occurring counts of seven textual relations through spoken-register CAs in each genre, with *n* referring to the raw occurring counts and its transformed counts per 1,000 units. Based on Table 18, in most genres, few textual relations are performed by spoken-register CAs except for *Listing* and *Contrastive*. Moreover, the occurrence of *Contrastive* is limited, with only one or two cases. It is *Listing* that appears most frequently in the spoken CA form. Due to the scarce occurrence of spoken-register CAs, statistical analysis is not employed in this part.

Table 18. Occurring counts of textual relations through spoken-register CAs across genres.

Textual Relation	Com-Con		Cau-Eff		Des		Def		Nar		Cla		Mul-Str		Arg		Pro-Sol	
	n	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000
Lis	2	17.5	5	37.9	4	34.2	3	24.4	3	17.0	0	0.0	6	35.7	5	23.5	6	25.7
Tra	0	0.0	0	0.0	2	17.1	0	0.0	2	11.3	0	0.0	0	0.0	1	4.7	0	0.0
App	0	0.0	0	0.0	0	0.0	1	8.1	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Sum	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	1	5.8	0	0.0	0	0.0	0	0.0
Res	1	8.8	1	7.6	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Con	1	8.8	0	0.0	1	8.6	3	24.4	2	11.3	0	0.0	0	0.0	1	4.7	1	4.3
Cor	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	3	17.3	0	0.0	0	0.0	0	0.0
Unit Count	114		132		117		123		177		173		168		213		232	

4.2 Distribution of Textual Relations Performed by CAs over Time

This part seeks to answer whether or not the distribution of CA-performed textual relations would vary in accordance to different time periods on a significant level. First, the results are presented without differentiating CAs in register. Then, the results are shown in terms of CAs in written register and in spoken register.

4.2.1 Distribution of Textual Relations Through All CAs over Time

To investigate the distribution of CA-performed textual relations over time, a temporal unit is set to be a semester. The data collected are sorted according to the semester when participants wrote the piece of writing. Table 19 presents the occurring counts of the seven textual relations performed by CAs in each of the four semesters. Again, the raw occurring counts are signaled by n, and transformed into the occurring counts per 1,000 units.

The results of examining the effects of textual relation and time on the distribution are attained via a two-way within-subjects ANOVA, with the two independent variables being textual relation and time as well as the dependent variable the occurring counts per 1,000 units. No interaction between textual relation and time ($F(18, 72)=0.912$, $p=0.567$) is found, nor is any main effect from time ($F(3, 12)=2.147$, $p=0.147$). Nevertheless, there exists a main effect from textual relation ($F(6, 24)=11.318$, $p<0.05$).

Table 19. Occurring counts of textual relations performed by all CAs over time.

Textual Relation	Time							
	Semester 2		Semester 3		Semester 4		Semester 5	
	n	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000
Lis	34	138.21	39	93.53	36	105.60	62	139.33
Tra	0	0	6	14.39	0	0	2	4.49
App	3	12.20	10	23.98	7	20.53	16	35.96
Sum	3	12.20	1	2.398	2	5.87	2	4.49
Res	7	28.46	8	19.18	10	29.33	26	58.43
Con	18	73.17	23	55.16	26	76.25	35	78.65
Cor	4	16.26	11	26.38	5	14.66	10	22.47
Unit Count	246		417		341		445	

Table 20 presents descriptive statistics of the occurring counts of CAs signaling 7 textual relations, and locates pairs of textual relations with significant differences. The results show that the distribution of CA-performed textual relations is the same as that found in the previous section. *Listing* and *Contrastive* occur the most frequently, *Summative* and *Transitional* appear the least frequently, and the occurring frequencies of *Resultive/Inferential*, *Appositive* and *Corroborative* rank in the middle. The results also show that while there is a significant difference among different compartments, no significant difference between textual relations within the same compartment is found.

Table 20. Significant differences between textual relations through all CAs over time and related descriptive statistics.

Lis	Tra	App	Sum	Res	Con	Cor		Mean	SD
---	0.019*	0.029*	0.022*	0.055	0.218	0.033*	Lis	23.832	6.405
0.019*	---	0.033*	0.676	0.003*	0.003*	0.028*	Tra	0.944	0.399
0.029*	0.033*	---	0.033*	0.136	0.017*	0.546	App	4.633	1.267
0.022*	0.676	0.033*	---	0.000*	0.001*	0.002*	Sum	1.248	0.589
0.055	0.003*	0.136	0.000*	---	0.010*	0.004*	Res	6.770	0.761
0.218	0.003*	0.017*	0.001*	0.010*	---	0.003*	Con	14.161	2.023
0.033*	0.028*	0.546	0.002*	0.004*	0.003*	---	Cor	3.989	0.839

4.2.2 Distribution of Textual Relations Through Written-Register CAs over Time

Table 21 presents the occurring counts of textual relations conveyed by written-register CAs in each semester. The raw occurring counts are symbolized by *n*, and then transformed into the occurring counts per 1,000 units.

After calculating a two-way within-subjects ANOVA, with the two independent variables being textual relation and time and the dependent variable the occurring counts per 1,000 units, it is found that no interaction between textual relation and time ($F(18, 72)=1.154$, $p=0.322$). However, there exist main effects from textual relation ($F(6, 24)=11.344$, $p<0.05$) and from time ($F(3, 12)=4.524$, $p<0.05$).

Table 22 presents related descriptive statistics and shows pairs of textual relations with significant differences in occurrence frequency. The results show that there exists a distribution norm, with *Listing* and *Contrastive* being the most frequent textual relations performed by written-register CAs. *Resultive/Inferential*, *Appositive,* and *Corroborative* are the second most. Finally, *Summative* and *Transitional* are the least frequent.

Table 23 presents related descriptive statistics and indicates the occurrence of significant difference among semesters in terms of written-register CA use. The results show that, except for semester 2, the use of written-register CAs significantly grows semester after semester.

Table 21. Occurring counts of textual relations through written-register CAs over time.

Textual Relations	Time							
	Semester 2		Semester 3		Semester 4		Semester 5	
	n	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000
Lis	26	105.69	28	67.10	30	88.00	52	116.90
Tra	0	0	2	4.80	0	0	1	2.25
App	3	12.20	9	21.60	7	20.53	16	35.96
Sum	1	4.07	1	2.40	1	2.93	0	0
Res	6	24.39	7	16.80	10	29.33	26	58.43
Con	17	69.11	20	48	26	76.25	33	74.16
Cor	4	16.26	11	26.40	2	5.87	10	22.47
Unit Count	246		417		341		445	

Table 22. Significant differences between textual relations through written-register CAs over time and related descriptive statistics.

Lis	Tra	App	Sum	Res	Con	Cor		Mean	SD
---	0.023*	0.033*	0.025*	0.074	0.355	0.040*	Lis	18.883	5.260
0.023*	---	0.036*	0.771	0.003*	0.002*	0.036*	Tra	0.352	0.233
0.033*	0.036*	---	0.028*	0.123	0.011*	0.373	App	4.513	1.244
0.025*	0.771	0.028*	---	0.123	0.011*	0.373	Sum	0.470	0.203
0.074	0.003*	0.123	0.001*	---	0.005*	0.001*	Res	6.446	0.844
0.355	0.002*	0.011*	0.002*	0.005*	---	0.001*	Con	13.374	1.774
0.040*	0.036*	0.373	0.026*	0.001*	0.001*	---	Cor	3.549	0.923

Table 23. Significant differences between semesters through written-register CAs over time and related descriptive statistics.

Semester 2	Semester 3	Semester 4	Semester 5		Mean	SD
---	0.290	0.817	0.124	Semester 2	6.620	1.226
0.290	---	0.390	0.007*	Semester 3	5.344	1.553
0.817	0.390	---	0.045*	Semester 4	6.368	0.959
0.124	0.007*	0.045*	---	Semester 5	8.860	1.270

4.2.3 Distribution of Textual Relations Through Spoken-Register CAs over Time

Table 24 presents the occurring counts of seven textual relations performed by spoken-register CAs in each semester, with *n* referring to the raw occurring counts and then transformed into the occurring counts per 1,000 units. Owing to the zero occurrences of many textual relations performed by spoken-register CAs, no statistical analysis is performed. Nevertheless, it is found that the use of spoken-register CAs and the types of textual relations performed by spoken-register CAs are diminishing in a steady fashion.

Table 24. Occurring counts of textual relations through spoken-register CAs over time.

Textual Relations	Time							
	Semester2		Semester3		Semester4		Semester5	
	n	*n per* 1,000	*n*	*n per* 1,000	*N*	*n per* 1,000	*n*	*n per* 1,000
Lis	8	32.520	11	26.379	6	17.595	0	0
Tra	0	0	4	9.592	0	0	1	2.247
App	0	0	1	2.398	0	0	0	0
Sum	2	8.130	0	0	1	2.932	2	4.494
Res	1	4.065	1	2.398	0	0	0	0
Con	1	4.065	3	7.194	0	0	2	4.494
Cor	0	0	0	0	3	8.798	0	0
Unit Count	246		417		341		445	

5. Discussion

This section respectively discusses the attained results concerning the distribution of CA-performed textual relations in the English writing by Chinese NS across genres and over time, and is wrapped up by a general discussion.

5.1 Discussion on Distribution of Textual Relations Through CAs Across Genres

A two-way ANOVA is calculated to examine the influence of genre and textual relation on the distribution of textual relations manifested by CAs. It is found that while genre has no role in impacting the distribution, there exists a norm distribution of textual relations across genres. In the norm distribution, textual relations are compartmented into three groups based on occurrence frequency. *Listing* and *Contrastive* occur most frequently. *Resultive/Inferential*, *Appositive*, and *Corroborative* have the second most occurrence frequency. *Summative* and

Transitional are the least frequent. The occurrence frequencies of the three groups significantly differ from one another, but for textual relations within any group, their occurrence frequencies do not differ from each other on a significant level. In addition, no significant difference among the three groups of textual relations is found when CA counts are differentiated by register. That is, the use of written-register CAs does not present a norm distribution of textual relations.

Table 25 presents a comparison between the norm distribution found in the present study and that in the previous studies. It is found that the two distributions highly resemble each other. Despite some differences in classification, *Listing* and *Contrastive* still occur most frequently, and *Transitional* is still the least frequent. Yet, *Resultive* and *Summative* in the present study occur less frequently than those in the previous studies.

Table 25. Comparison between the distributions of the textual relations.

Distribution	Most frequent		Moderate frequent		Least frequent
Distribution of the present study	Listing Contrastive	>	Resultive/Inferential Appositive Corroborative	>	Summative Transitional
Distribution of the previous studies*	Listing Contrastive Resultive	>	Appositive Summative	>	Transitive Inferential

*The distribution is based on Table xx in Section 2.2.

The lack of genre influence might be in relation to textual-relation-creating mechanism. Textual relations can be manifested by various mechanisms. For instance, CAs can be easily replaced with and rephrased by discourse-organizing words (McCarthy, 1991; Winter, 1977; Yu, 2007). As shown in the following example, the CA in (1) can be rephrased as the noun phrase in (2), and the listing relation is still conveyed. Thus, while various genres may have their own distinct discoursal characteristics, the CA use alone may not be sufficient to distinguish genres.

Taiwan officially becomes an aged society.

1) Firstly, the birth rate is substantially declining.

2) The first reason is that the birth rate is substantially declining.

Another reason to explain the lack of genre influence is that genres are not mutually exclusive. Even though different genres may be constructed with different discoursal characteristics, they may also incorporate characteristics from one another, especially when writing length becomes longer. Considering that some of the writing pieces are twice the length of the others in the present study, the nine genres investigated may share many characteristics, which, in some sense, makes the nine genres a general superordinate genre. Therefore, no genre influence is found.

As opposed to lack of genre influence, CA-performed textual relations present a constant distribution norm across genres. Two explanations are proposed to account for it. The first explanation is that the nature of different textual relations has forecast their occurrence frequencies. For example, it is understandable that *Summative* is in the group where textual relations occur least frequently, because *Summative* indicates a conclusion which only appears at the end of a text no matter how long the text is. In contrast, *Listing* occurs most frequently, because writers can always employ *Listing* CAs to indicate more ideas to come without limitation.

The second explanation for a norm distribution of textual relations is that there exists a preference programmed in human cognition for employing CAs to convey certain textual relations. Take *Contrastive* as example. The textual relation is relatively complicated because it requires an effort to analyze two events and to locate the contrastive points. Therefore, it would take more energy to describe the relation in text compared with writing in the common temporal sequence. Due to the extra energy required, Economy Principle is applied in order to minimize the energy consumption while to successfully achieve communication (Ungerer & Schmid, 2006). Based on the rationale, it is inferred that the tendency to select CAs, rather than other textual-relation-creating devices, to convey the contrastive relation is programmed in human cognition since one or two words of CAs are enough to minimize the energy consumption and to express the textual relation clearly. Ultimately, *Contrastive* becomes one of the textual relations most frequently performed by CAs.

Two pieces of evidence may support the preset preference of certain textual relations manifested by CAs in human cognition. The observational evidence is that regardless of genre difference, *Contrastive* has the highest occurrence frequency in all NS data and some NNS data from previous studies (Chen, 2006; Altenberg & Tapper, 1998; Field & Yip, 1992; Shen, 2006) reviewed in section 2. The statistic evidence is that a distribution norm of textual relations is found through all CAs, but not through written-register CAs. Because register is a manmade literary concept, neither written-register CAs nor spoken-register ones can completely reflect human cognition. Therefore, deliberately exploring written-register CAs alone and ignoring spoken-register CAs fails to construct a distribution norm of textual relations on a significant level. Only when all CAs are considered without register differentiation can human cognition be fully represented by a norm distribution of textual relations across genres.

5.2 Discussion on Distribution of Textual Relations Through CAs over Time

To explore whether time and textual relation affect the distribution of textual relations conveyed by CAs, a two-way ANOVA is calculated, and the results are similar to those found when exploring the effects of genre and textual relation on the distribution. A distribution norm of CA-performed textual relations free from time influence is found. The distribution norm is divided into three groups. *Listing* and *Contrastive* are the most frequent textual relations manifested by CAs, *Resultive/Inferential*, *Appositive* and *Corroborative* are the second most frequent, and *Summative* and *Transitional* are the least frequent. The three groups are significantly different from one another in terms of occurrence frequency, and textual relations within any group do not. However, when CAs are separated from written register from spoken register, time is found to have a main effect on the CA use, with more use of written-register CAs in later semesters.

The lack of time influence on the distribution of textual relations might be in relation to cognitive development. According to Inhelder and Piaget (1999), the development of logical thinking reaches maturation after adolescence. Therefore, it may be assumed that once the logical thinking becomes less variable, an innate distribution norm of textual relations to express ideas in human cognition may emerge accordingly. Since the data sources in the present study are college students with mature cognition, time is no longer a factor affecting their use of CAs to perform various textual relations. Instead, a distribution preference is reflected when textual relations are performed by CA.

In contrast, according to the statistic results, time has a main effect on the register use of CAs. Over time, the use of written-register CAs is significantly increasing while that of spoken-register CAs is decreasing. The result is understandable. Since register is often taught and then acquired by writing learners through education, the more time students stay in school and receive writing training, the more skillful students are to write in written register and avoid spoken register. Moreover, the fact that register use can be taught over time while the distribution of textual relations via CAs is immune to time highlights the possibility of a norm distribution existing in human cognition, because the norm distribution cannot be taught and changed over time.

5.3 General Discussion

According to the results, neither genre nor time has an effect on the occurrence frequencies of CAs manifesting various textual relations. Instead, a distribution norm of CA-performed textual relations across genres and over time is found. In the distribution norm, *Listing* and *Contrastive* have the highest occurrence frequency. *Summative* and *Transitional* have the least. *Resultive/Inferential*, *Appositive* and *Corroborative* rank in the middle of the frequency order.

As reviewed in section 2, previous studies also reported similar results. A distribution independent of genre and time was found in the NS writing in studies based on four CA-performed textual relations (Chen, 2006; Field & Yip, 1992) as well as in both NS and NNS writing in studies based on more CA-performed textual relations (Altenberg & Tapper, 1998; Shen, 2006; Tankó, 2004). Moreover, the formations of these found distribution patterns are very similar as well. *Contrastive* which is based on the more-type framework and equates *Adversative* in the four-type framework, usually occurs most frequently. *Summative* and *Transitional* happen least frequently.

The striking similarity between the results in the present study and those in previous studies is in support of the contention that there exists a preset distribution preference of using CAs to manifest various textual relations in human cognition. It reflects how the human mind perceives these textual relations in terms of logical complex. In other words, the distribution norm of CA-performed textual relations based on CA occurrence frequency is a mental representation of human cognition. The register factor provides more evidence to support the contention. When CAs are divided by register, no distribution pattern of textual relations performed by written-register CAs can reach a significant level. This is because, without elements embodied by spoken-register CAs in real world, the mental representation becomes flawed, and it is this incomplete mental representation that no distribution pattern exactly reflects.

Since the data source in the present study is the English writing by Chinese NS, another relevant issue is whether the found distribution norm also underlies the English writing by its NS. In light of the fact that the found distribution norm in the present study is very similar to the distribution identified in both NS and NNS writing in studies with a framework of fine classification (Altenberg & Tapper, 1998; Shen, 2006; Tankó, 2004), it is suggested that the found distribution pattern may be insusceptible to the first language influence. That is to say, the found distribution is universal in the English writing in spite of writers' language background. Any English writing pieces where the found distribution cannot be extracted may be regarded as ill-composed whether the writers are English NS or NNS.

The finding has great application potential in automation of discourse diagnosis. Up to date, researchers has attempted to develop automatic tools to diagnose English writing on a discourse level based on sentence length, syllable counts, and difficulty levels of vocabulary (Chall & Dale, 1995; Klare, 1984). The outcome has not been satisfactory for these linguistic characteristics are on local and shallow levels (Bailin & Grafstein, 2001). Benjamin (2012) points out that only by taking into account factors on global and deeper levels can automatic tools judge whether a writing piece constructs a coherence mental representation and produce reliable discourse diagnosis. On that note, the distribution found in the present study can serve as a crucial criterion for automatic tool development. Such a possible tool can extract CAs in a piece of writing and compare the distribution pattern of textual relations performed by these extracted CAs to the found distribution norm. Based on the matching degree, how well-constructed the piece is can be evaluated accordingly. What's better, since the found distribution is not subject to genre, time, and the first language influence, tools featuring the distribution criterion can be available to an all-inclusive variety of users.

6. Conclusion

The present study began with an investigation into the use of CAs performing various textual relations, and discovered that a distribution norm of CA-performed textual relations based on CA occurrence frequency persists across genres and over time. The found distribution can serve as an indicator of discoursal coherence. For a piece of English writing presenting the found distribution during discourse analysis, it may be considered potentially coherent. Instructors can also point out the incoherence in learners' writing by referring to the deviation from the found distribution. The study ended up suggesting using the found distribution as an evaluating criterion for developing automatic tools of discourse diagnosis.

For further research, two possibilities await. Firstly, *Listing* and *Contrastive* are two textual relations with highest occurring frequencies. While the high frequency of *Contrastive* can be explained by the cognitively economic reason, that of *Listing* is said be rooted in teaching instructions (Shen, 2006). Due to the insufficient English proficiency of NNS, NNS might be encouraged to employ more *Listing* CAs because it is quick to construct the textual structure, which leads to *Listing* ubiquity in the NNS writing. Thus, even though genre and time have no influence on the found distribution, whether teaching instructions plays a role in the distribution remains unknown. Secondly, the mechanisms to realize textual relations are not limited to the CA use. Whether the found distribution still holds after including the derived and paraphrased forms of CAs, such as preposition expressions with references and discourse-organizing words, is also worth pursuing.

References

Altenberg, B. & Tapper, M. (1998). *The use of adverbial connectors in advanced Swedish learners' written English.* (First edition ed.). In S. Granger (Ed.), Learner English on Computer (pp. 80-93). Longman.

Atkins, S., Clear, J. & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*, *7*(1), 1-16. https://doi.org/10.1093/llc/7.1.1

Bailin, A., & Grafstein, A. (2001). The linguistic assumptions underlying readability formula: A critique. *Language & Communication*, *21*(3), 285-301. https://doi.org/10.1016/S0271-5309(01)00005-2

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, *24*, 63-88. https://doi.org/10.1007/s10648-011-9181-8

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Pearson Education.

Bolton, K., Nelson, G. & Hung, J. (2002). A corpus-based study of connectors in student writing: Research from the International Corpus of English in Hong Kong (ICE-HK). *International Journal of Corpus Linguistics*, *7*(2), 165-182. http://dx.doi.org/10.1075/ijcl.7.2.02bol

Celce-Murcia, M. & Larsen-Freeman, D. (1999). *The Grammar Book: An ESL/EFL Teacher's Course*. (Second Edition ed.). Heinle & Heinle Publishers.

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

Chen, C. W.-y. (2006). The use of conjunctive adverbials in the academic papers of advanced Taiwanese EFL learners. *International Journal of Corpus Linguistics*, *11*(1), 113-130. http://dx.doi.org/10.1075/ijcl.11.1.05che

Connelly, M. (2013). *Get writing: Sentences and paragraphs*. Thomson Higher Education.

Field, Y. & Yip, L. (1992). A comparison of Internal conjunctive cohesion in the English essay writing of Cantonese speakers and native speakers of English. *RELC Journal*, *23*(1), 15-28. https://doi.org/10.1177/003368829202300102

Granger, S. & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, *15*(1), 17-27. https://doi.org/10.1111/j.1467-971X.1996.tb00089.x

Halliday, M.A.K. & Hasan, R. (1976). *Cohesion in English*. Routledge.

Huang, K.-H. (2018). *Textual Cohesion and Cohesive Errors in English Writing of Taiwan University Students* (Master’s thesis). Retrieved from https://hdl.handle.net/11296/35mj85

Inhelder, B. & Piaget, J. (1999). *The Growth of Logical Thinking from Childhood to Adolescence*. Routledge.

Klare, G. (1984). Readability. In Pearson, P. *et al.* (Eds), *Handbook of Reading Research*, (pp. 681-744). Longman.

Langan, J. (2010). *Paragraph Skills*. McGraw-Hill.

Lannon, J. M. (2007). *The writing process: A concise rhetoric, reader, and handbook.* Longman.

Liu, M. & Braine, G. (2005) Cohesive features in argumentative writing produced by Chinese undergraduates. *System*, *33*(4), 623-636. https://doi.org/10.1016/j.system.2005.02.002

McCarthy, M. (1991). *Discourse Analysis for Language Teachers*. Cambridge University Press.

Ministry of Education (MOE). (2014). *Curriculum Guidelines of 12-Year Basic Education: General Guidelines*. Retrieved May 03, 2020, from https://www.naer.edu.tw/files/15-1000-14113,c639-1.php?Lang=zh-tw

Ministry of Education (MOE). (2018). *Curriculum Guidelines of 12-Year Basic Education: Guidelines for English.* Retrieved May 03, 2020, from https://www.naer.edu.tw/files/15-1000-14113,c639-1.php?Lang=zh-tw

Morenberg, M. & Sommers, J. (2008). *The writer's options: Lessons in style and arrangement*. Pearson Longman.

Morey, M., Muller, P. & Asher, N. (2018). A Dependency Perspective on RST Discourse Parsing and Evaluation. *Computational Linguistics*, *44*(2), 197-235. https://doi.org/10.1162/COLI_a_00314

Phoocharoensil, S. (2017). Corpus-based exploration of linking adverbials of result: Discovering what ELT writing coursebooks lack. *The Southeast Asian Journal of English Language Studies*, *23*(1), 150-167. http://doi.org/10.17576/3L-2017-2301-11

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman.

Reid, J. M. (2000). *The process of composition*. Longman.

Shen, T.-C. (2006). *Advanced EFL Learners' Use of Conjunctive Adverbials in Academic Writing* (Master's Thesis). Taiwan: National Taiwan Normal University.

Simons, G. & Bird, S. (2008, May 31). *OLAC Metadata. OLAC: Open Language Archives Community*. Retrieved May 14, 2020, from http://www.language-archives.org/OLAC/metadata.html

Smalley, R. L., Ruetten, M. K. & Kozyrev, J. R. (2012). *Refining composition skills: Academic writing and grammar*. Cengage.

Suehring, S. (2006). Beginning web development with Perl: from novice to professional. Springer.

Tankó, G. (2004). *The use of adverbial connectors in Hungarian university students' argumentative essays*. In Sinclair, J. M. (Ed.), How to Use Corpora in Language Teaching (pp. 157-181). John Benjamins. https://doi.org/10.1075/scl.12.13tan

Tseng, Y.-C. & Liou, H.-C. (2006). The effects of online conjunction materials on college EFL students' writing. *System*, *34*(2), 270-283. https://doi.org/10.1016/j.system.2006.01.006

Ungerer, F. & Schmid, H.-J. (2006). *An Introduction to Cognitive Linguistics*. Routledge.

Winter, E. O. (1977). A clause-relational approach to Eenglish texts: A study of some predictive lexical items in written discourse. *Instructional Science*, 6, 1-92. https://doi.org/10.1007/BF00125597

Wyrick, J. (2008). *Steps to writing well: With additional readings*. Wadsworth.

Yu, H.-y. (2007). Discourse grammar for academic reading: Textual relationships. *English Teaching & Learning*, *31*(2), 159-197. https://doi.org/10.6330/ETL.2007.31.2.05

Xie, J.-Z. (2014). *Textual Cohesion in Senior High School Students' Expository Writings in Southern Taiwan* (Master's thesis). Taiwan, National Pingtung University of Education.

Computational Linguistics and Chinese Language Processing
Vol. 26, No. 1, June 2021, pp. 79-104

中文新聞文本之宣傳手法標記與分析

The Analysis and Annotation of Propaganda Techniques in Chinese News Texts

施孟賢[*]、段人鳳[+]、鍾曉芳[#]

Meng-Hsien Shih, Ren-feng Duann, Siaw-Fong Chung

摘要

新聞媒體常在政治新聞文本中運用宣傳手法(propaganda techniques)表達媒體本身之政治立場，企圖影響讀者之立場。目前尚無具宣傳手法標記之中文語料供立場分析，本文以可解釋性的方式，人工細部標記中文新聞文本所使用之宣傳手法、並以 Bootstrap 方式擴展標記規模的資料集，再分別以人工檢核與先導實驗來確保標記資料集之效能。透過單純貝式分類器搭配基本的詞袋特徵進行訓練後，機器判讀行段是否包含宣傳手法的準確率達 74.26%。本宣傳手法之人工標記資料已公開釋出，可應用於未來機器訓練與學習預測新文本之立場。

Abstract

In political news media, propaganda techniques are often employed to express one's political view, or to influence the audience's stance. Chinese corpora with the

[*] 國立清華大學臺灣語言研究與教學研究所；國立中正大學通識教育中心
Institute of Taiwan Languages and Language Teaching, National Tsing Hua University; Center for General Education, National Chung Cheng University
E-mail: simon.xian@gmail.com

[+] 國立臺東大學通識教育中心（通訊作者）
Address for correspondence: Center for General Education, National Taitung University. No. 369, Sec. 2, University Road, Taitung, Taiwan 950, Republic of China
E-mail: rduann@nttu.edu.tw. Tel: 886-(0)89-517955. Fax: 886-(0)89-517494

[#] 國立政治大學英國語文學系
Department of English, National Chengchi University
E-mail: sfchung@nccu.edu.tw

annotation of propaganda techniques are yet to be developed. In this paper, with an explainable approach, we annotated the use of propaganda techniques in Chinese political news texts, and enlarged the dataset by bootstrapping using a small set of manually annotated data. To ensure the validity, we manually corrected the bootstrapped dataset and ran a pilot machine-learning experiment using a naïve Bayes classifier trained with the bag-of-words feature. A precision of 74.26% was reached for the binary classification (with or without propaganda technique). The manually annotated data with propaganda techniques is available online for the application of machine training and learning to predict the stance of new texts.

關鍵詞：情感（立場）分析、語言資源、宣傳手法、台灣新聞媒體

Keywords: Sentiment (Stance) Analysis, Language Resource, Propaganda Techniques, Taiwan News Media

1. 緒論 (Introduction)

宣傳(propaganda)一詞[1]，原為宗教用語，表示羅馬天主教會的傳教活動，於 19 世紀中葉該詞漸漸用於政治領域，且開始帶有貶意(Diggs-Brown, 2011)。Lasswell (1927, p. 9)將宣傳定義為「使用重要的符號來控制意見」[2]。大英百科全書也將之定義為「訊息（事實、論點、謠言、半事實、或謊言）的傳播，以期影響輿論」(Smith, 2020)。Zollman (2019, p. 335)引述 Bussemer (2005, pp. 29-30)，將宣傳定義為「透過象徵性的傳播，散布經由媒體所形成之政治或社會團體對某事件的見解／態度」以及「產製支持特定利益的群眾」的行為。報紙做為宣傳的平台由來已久，Riegel (1935, p. 206)即指出，單就新聞選材而言，報紙即遵循一套本質為政治的準則。而美國人普遍認為報紙充斥商業與政治宣傳，因而對報紙內容始終存疑。Riegel 確立了報紙作為一個宣傳場域的事實。另一方面，該論文指出當代對於商業與政治宣傳的檢視方法，大多基於主觀、短暫並且不確定的參考方式(p. 202)。過往研究試圖提出一套準則，俾使研究者檢視宣傳的技巧（如 Lee & Lee, 1939; Weston, 2018）。近年來，隨著科技的發展及語料蒐集技術的演進，計算學者提出新方法，以客觀、一致的準則來偵測報紙或新聞媒體中的宣傳手法（如 Barrón-Cedeño *et al.*, 2019; Da San Martino *et al.*, 2019）。然而，這些研究都是以英文新聞媒體為探討對象，針對中文新聞媒體的宣傳手法辨識的方法則付之闕如。

本研究以 Da San Martino *et al.* (2019)所提出的 18 種宣傳手法為出發點，分析台灣兩個政治立場相反的報紙對同一事件撰寫的社論，提出中文報紙使用的宣傳手法，為第

[1]Diggs-Brown (2011)指出，propaganda 衍生自 1622 年由天主教會所創的 Congregatio de propaganda fide（意指由樞機主教組成的行政單位，其任務為在天主教國家執行教會事務）。1790 年代，propaganda 的英文擴及非宗教層面，19 世紀中葉該詞漸漸用於政治領域，且開始帶有貶意。本文將 propaganda 翻譯為「宣傳」。

[2] 本文引述的英文文獻皆由作者翻譯為中文。

一篇以語料庫探究、提取中文新聞媒體宣傳手法的研究。本研究所辨識的方法與詞彙，可做為日後自動偵測的訓練語料。研究發現，由於報紙社論的特質，Da San Martino *et al.* (2019)的諸多方法並未出現在本研究語料中。再者，我們也發現在 Da San Martino *et al.* 提出的手法之外，中文報紙尚採用了兩個手法：(1) 引用歷史：報紙引用歷史事件或人物，讓讀者連結、類比當前事件與歷史事件，意圖影響讀者立場；(2) 預設立場：報紙透過修辭性問句(rhetorical question)和評價標記(evaluative marker)來表明其預設立場，藉以暗示並影響讀者對該事件的看法。就中文報紙使用的部分宣傳手法，本研究也給予更明確的語言上的定義。同時，某些手法出現次類，我們也針對這些次類舉例說明。

本文主要回答下列三個研究問題：

一、 中文政治新聞文本中使用的宣傳手法為何？

二、 中、英文政治新聞採用的宣傳手法有何異同？

三、 以 Bootstrap 方式擴增中文政治新聞宣傳標記之成效如何？

本文第 2 章回顧新聞媒體與宣傳、篇章分析與宣傳，以及計算機語言學之相關研究；第 3 章說明本研究採用的語料、標記方法、檢核方式與先導實驗；第 4 章陳述結果與討論，第 5 章則說明本研究的貢獻與限制。

2. 文獻回顧 (Literature Review)

立場偵測(stance detection)與情感分析(sentiment analysis)是相關但不同的工作，情感分析通常著重於決定文本為正面、負面或中性，立場偵測則在於確認文本傾向的特定對象，亦即「立場」。而相同的立場可能經由正面的語言表達，也可能以負面的語言表達(Mohammad *et al.,* 2017)，並無法直接藉由文本的情感分析獲知其立場，因此立場偵測有其獨特重要性，且廣泛運用於資訊擷取(information retrieval)與文本摘要(text summarization)領域。

新聞媒體的立場對大眾有一定的影響力，且常藉由宣傳手法(propaganda techniques)表達其立場，進而影響讀者對新聞事件之看法。Zollmann (2019, p. 329)認為，「對於形成輿論而言，宣傳舉足輕重」，並且「新聞媒體顯然成為擴散宣傳的管道」。然而，Zollmann 也指出，「當代傳播、媒體和新聞等相關研究多半忽略了媒體在製造與散布宣傳時所扮演的角色」。

Zollman (2019, pp. 337-341)亦參考了 Kromrey (2009)對於宣傳面向的看法，進一步說明了宣傳的三個面向：

面向 1：整合新聞媒體宣傳與意識形態（透過省略某些觀點並以框架呈現某事件、議題或參與者；刪除重要的評論；使用在意識形態上具爭議性的概念，例如「戰爭」、「犯罪」、「屠殺」、「種族滅絕」、「恐怖主義」、「民主」、或「社會主義」等）(Herman & Peterson, 2010; Keeble, 1997; Zollmann 2017，轉引自 Zollmann, 2019, p. 337)。

面向 2：有技巧的新聞媒體宣傳與「事實」（扭曲事實，邊緣化事實或敘述）(Zollmann, 2019,

p. 340)。

面向 3：妖魔化(demonising)新聞媒體宣傳與惡行（負面聯想、貼標籤）(Zollmann, 2019, p. 341)。

以篇章分析(discourse analysis)探討新聞的研究在西方始於 1970 年代後期(Hodge, 1979)，且持續獲得學者重視，例如 Fowler (1991)、Fairclough (1995)、van Dijk (1988, 1995)、van Leeuwen (2008)…等，皆針對新聞中的用字遣詞與再現，來揭露新聞欲傳達的意識形態。在中、英新聞報導的篇章分析對比研究方面，Wang (2007)探討了中文和（澳洲）英文報紙對 911 事件評論中的文本特徵和互文使用，洪見合(2014)則分析中文與英文報紙社論的篇章修辭架構和用字特點，兩者皆聚焦不同語言和文章結構之互動，並未探討新聞中的宣傳。新聞宣傳之篇章分析相關的研究屈指可數，Patrona (2018)探討了希臘境內歷時 15 年主流黃金時段的新聞，發現希臘的新聞界出現了「個人意見／觀點升級為『事實』的新聞報導」的現象(p. 185)，再者，新聞已轉變為帶有政治宣傳的「對話」，並具有形塑輿論的功能。【例 1】摘自 Patrona (2018, pp. 190-191)分析 2015 年 6 月 28 日 SKAI 頻道的新聞(H 代表女主播，P 代表記者)：

【例 1】

```
H the prime minister used the word shame, addressing
  it to the European institutions
  (0.1)
P
  ((gesturing)) maybe - but here it is clear, (.)and
  unfortunately, (.) we must see it as it is -
  ((raised pitch)) it's being revealed before our
  eyes (.) and we will not be able to say on Sunday
  night, we didn't see it - it's being revealed, that
  Mr Tsipras has a plan which he is keeping ((raised
  pitch)) secret
  ((turns omitted))
```

【例 1】中，Patrona 說明畫底線的詞語‘institutions’、‘unfortunately’、‘it's being revealed’、‘our eyes’、‘secret’等，皆帶有政治宣傳意味；再者，這些詞語以不同形式呈現，可以是單詞、短語、子句等。換句話說，宣傳所使用的語言不限於單一形式。

Bensa & Wijaya (2017)探討了 2016 年中國主張的九段線(nine-dashed line，9DL)納入了印尼的那吐納水域(Natuna waters)，印尼向中國駐印尼使館口頭抗議後，印尼的 Kompas 和中國的新華社報導中，中國和印尼的關係。Bensa & Wijaya 分析印尼媒體採用的宣傳手法，包含「貼標籤」、「裝飾法」、「轉移法」、「親民法」、「證詞法」、「選擇法」、「從眾效應」及「替罪羔羊」等，表 1 為該論文對這些方法的定義與範例：

表 1. 新聞媒體 8 種宣傳手法(Bensa & Wijaya, 2017, p. 3)
[Table 1. Eight media propaganda techniques (Bensa & Wijaya, 2017, p. 3)]

序號	手法	定義	範例
1.	Name Calling（貼標籤）	讓觀眾不經思考，直接接收宣傳者觀點的手法。通常使觀眾以負面概念去理解某個人物或想法。	宣傳者稱呼其政敵為「陰謀者」、「騙子」、「恐怖份子」
2.	Glittering Generality（裝飾法）	使用正向評價的概念與信念，抓住觀眾的情緒，確保目標觀眾感覺正向、積極，並進而接受宣傳者要傳達的內容。	印尼是個民主國家！
3.	Transfer（轉移法）	將欲宣傳的事物和受尊敬的象徵連結起來，以改善該事物的形象。這個手法可促使觀眾認同特定權威。	使用代表輿論共識的山姆大叔。
4.	Testimonial（證詞法）	請名人（演員／政治人物）宣傳商品或為候選人站台。這一手法訴求觀眾的情緒而非邏輯思考，因為證詞內容的邏輯可能相當薄弱。	歐普拉支持歐巴馬競選總統。
5.	Plain Folks（親民法）	將宣傳者的形象塑造為平民百姓，讓觀眾認為宣傳者是自己的一份子，並相信宣傳者與民同在、苦民所苦。	Maspion 為印尼製造，請使用印尼製產品。
6.	Card Stacking/ Selection（選擇法）	選擇片面事實來支持宣傳者欲保障的利益，或證明其以往的言論。	自由市場的負面效應。
7.	Bandwagon（從眾效應）	心理學說明了個人內心都想和大多數人一樣，宣傳者於是利用這個心態，暗示觀眾倘若不和其他人一樣，將會錯過良機甚或有所損失。	「每個人都這麼做，你也該這麼做。」
8.	Frustration of Scapegoat（代罪羔羊）	當問題發生時，宣傳者找尋代罪羔羊以便將所有問題歸咎於他／她。這個手法也可以引發觀眾的憤恨。	「都是他／她(害的)。」

Bensa & Wijaya 雖探討了媒體的宣傳手法， 卻未說明如何辨識這些宣傳手法，在相關的研究方法上，仍有待後續研究發展。

計算機語言學對宣傳的探討，早期以英文為研究主體的文獻中，大多從文件的層次 (document level) 偵測宣傳文章，甚至僅標記文章出處是否為宣傳來源，而將該來源所有

文章視為宣傳。但 Horne *et al*. (2018)指出，宣傳的文章來源也會定期發佈客觀的非宣傳文章以提高該來源之可信度。此外，以文件層次標記訓練出來之計算模型較缺乏可解釋性(explainability)。

另一方面，Barrón-Cedeño *et al*. (2019) 提出一套以新聞寫作風格、可讀性和詞彙豐富性為基礎的模型，試圖偵測新聞報導是否隱含宣傳目的，並計算其宣傳分數(propaganda score)。該研究結果顯示，這套系統優於過往的偵測系統，但該文也建議相關學者未來可研究以語句的片段層次(fragment level)來辨識新聞文本宣傳手法。有鑑於此，Da San Martino *et al*. (2019, pp. 5637-5639)聚焦於語句的片段層次，針對來自宣傳性新聞媒體網站的文章中，可從內部判定、無須外在訊息支援之宣傳文章，列出表 2 的 18 種宣傳手法，並據此檢視 372 篇英文新聞（約 35 萬詞）。

表 2. 英文新聞的 18 種宣傳手法(Da San Martino et al., 2019, pp. 5637-5639)
[Table 2. Eighteen propaganda techniques in English news articles (Da San Martino et al., 2019, pp. 5637-5639)]

1. Loaded language[3] **2. Name calling** 3. Repetition **4. Exaggeration or minimization** **5. Doubt** 6. **Appeal to fear**	**7. Flag-waving** 8. Causal oversimplification 9. Slogans **10. Appeal to authority** **11. Black-and-white fallacy** **12. Thought-terminating** *cliche*	13. Whataboutism 14. Reductio ad Hitlerum 15. Red herring 16. Bandwagon 17. Obfuscation 18. Straw man

本研究以這 18 種手法為出發點，並參考 Da San Martino *et al*. (2020) 新標記之 550 篇新聞文本資料[4]，以分析中文報紙社論的宣傳手法。表 2 列出的 18 個英文新聞宣傳手法之定義與實例詳見「3. 研究方法」中的表 4。

以中文為主體的相關研究多半聚焦中國的社群媒體檢查制度。Knockel *et al.* (2015) 探討中國如何執行資訊管控，該研究針對中國人常用的四個社群影音平台，透過逆向工程(reverse engineering)，辨識客戶端的關鍵詞檢查制度，並歸納出分屬六大主題（政治、社會、人物、事件、科技、雜項）共計 17,547 關鍵詞詞表，該研究發現，不同平台對執行審查仍留有各自的彈性，再者，與「集體行動」和「批評政府」相關的貼文常為審查標的。Arefi *et al.* (2019) 沿用了 Knockel *et al.* (2015)的詞表，採用深度學習、卷積神經網路定位(CNN localization)及自然語言處理技術(NLP techniques)，針對 14 個受審查的類別[5]，探討中國新浪微博受到審查和未受審查的貼文與圖片。該研究發現，情感(sentiment)

---

[3] 表 2 以粗體字呈現之宣傳手法亦出現於本研究的中文新聞文本中。

[4] 標記新聞資料取得方法參考此網頁 http://propaganda.qcri.org/semeval2020-task11。

[5] 這 14 個類別為：(1) 薄熙來、(2) 鄧小平、(3) 火災、(4) 死亡／傷害、(5) 劉曉波、(6) 毛澤東、(7) 人民大會、(8) 軍／警、(9) 抗爭、(10) 色情／裸露、(11) 暴雨、(12) 小熊維尼、(13) 習近平、(14) 周克華。

為不同主題的唯一共通審查指標，不同主題也受到不同層次的審查，而受到審查的貼文僅有少於三小時的網路存活時間。上述研究所使用的關鍵詞，皆以社群平台貼文主題為分類基準，以宣傳手法作為詞彙／語句片段層次分類基準的研究仍有待進一步發展。

在標記品質的評估測量上，有別於傳統的分類標記問題，宣傳手法之標記屬於序列標籤(sequence labeling)問題，需同時考慮標記不定長度之單詞／短語／子句，以及該單詞／短語／子句之標籤，另需考量不同標記者之標記文字部份重疊之評測。有鑑於此，Mathet, Widlocher & Métivier (2015, pp. 441-447)提出 γ 評測以評估此類序列標籤的標記者間一致性。惟本文做為首篇標記中文宣傳手法之研究，目前僅初步評測標記者間針對同一行的標記是否一致的百分比(Scott, 1955, p. 323; Artstein & Poesio, 2008, p. 558)，詳細的評測基準請見本文「4.結果與討論」。

3. 研究方法 (Methodology)

本文選擇的中文報紙為台灣兩份政治立場相反的主流報紙：自由時報與聯合報。自由時報傾向由民主進步黨領導的陣營，在國家認同上主張台灣獨立，因而對中國及其相關事務採取保留的態度；聯合報則支持由中國國民黨領導的陣營，認同 1912 年建立的中華民國，對中國及相關議題較為友善。

本研究選出上述兩報有關太陽花學運的社論。選擇社論的原因，乃根據 Smith (2020)說明，社論可視為隱蔽型(covert)的宣傳，即閱聽人對宣傳者身份或來源一無所知。Riegel (1935)不諱言社論具有宣傳的特質，以往亦有社論作為宣傳的研究（例如，賴瞖，1965；游承季，2001），我們因而將報紙社論視為宣傳的一類。

太陽花學運為一場由大學生主導的社會運動，肇因於當時的執政黨（中國國民黨）在立法院企圖迅速通過海峽兩岸服務貿易協定（簡稱「服貿」），這項協定被抗議學生與公民團體認為有損台灣現狀並會危及台灣中小企業，他們因而進入立法院，佔領立法院長達 24 天之久(2014 年 3 月 18 日至 4 月 10 日，Lin & Hsieh, 2017)，創下台灣歷史上立法院被佔領的首例。

本研究的新聞資料來自台灣的自由時報與聯合報網站，搜尋日期介於 2014 年 3 月 18 日至 2016 年 12 月 31 日之間，含「服貿」或「太陽花」關鍵詞之社論。如表 3 所示，自由時報有 78 篇（佔本語料庫文章數的 34%）共 147,562 詞（佔本語料庫詞數的 41%），分為 17,098 行；聯合報有 150 篇（佔本語料庫文章數的 66%）共 213,437 詞，（佔本語料庫詞數的 59%），分為 17,888 行。經過初步標記兩份報紙日期最早的前 10%社論（自由時報自 2014 年 3 月 18 日至 2015 年 3 月 20 日共 8 篇 1,073 行，聯合報自 2014 年 3 月 18 日至 2014 年 4 月 5 日共 15 篇 1,825 行）的宣傳手法之後，我們發現，Da San Martino *et al*. (2019)提出的用於標記英文新聞的18種宣傳手法並未完全出現於中文新聞文本上。由於本研究採用的語料為報紙社論，而社論主要為「報社對新聞議題深思熟慮後所形成之見解，此見解以肅穆、說理的筆觸，常態地呈現在每期報刊上，目的在於聲明報社對該議題之立場，並欲勸服讀者採取一致的立場。」（朱灼文, 2003, pp. 18-19）；社論可

用來倡導或反對某主張，內容多半採取「理性客觀的口吻、莊重持平的語調」(p. 358)，因此，諸如過度簡化因果(Causal oversimplification)、「那又怎麼說」主義(Whataboutism)、刻意模糊與混淆(Obfuscation, intentional vagueness, confusion)、轉移焦點(Red herrings)...等缺乏理性論證與完整討論，或不合邏輯的手法便不會出現在本研究的語料中。

表 3. 本研究從自由時報和聯合報取得太陽花學運相關社論文章的統計
[Table 3. Basic information of the editorials of the Liberty Times and the United Daily News]

	自由時報	聯合報	總和
文章數	78 (34%)	150 (66%)	228 (100%)
詞數	147,562 (41%)	213,437 (59%)	360,999 (100%)
行數	17,098 (49%)	17,888 (51%)	34,986 (100%)

再者，我們在標註語料時，也發現某些詞語無法歸類於 Da San Martino *et al*. 提出的 18 項手法，我們認為這是中文新聞獨有的宣傳手法，因此將之獨立出來，並分別命名為「引用歷史」和「預設立場」，茲說明如下：(1) 引用歷史：報紙引用歷史事件（例如，美麗島事件、天安門事件）或人物，語言特徵為「『台（灣）版』＋歷史事件名稱」或「歷史事件名稱＋『翻版』」，或使用類比或比較文字「如」、「像」 、「相較」，讓讀者連結、類比當前事件與歷史事件，意圖影響讀者立場；(2) 預設立場：報紙透過修辭性問句(rhetorical question)和評價標記(evaluative marker)來表明其預設立場，藉以暗示並影響讀者對該事件的看法。關於修辭性問句，根據湯廷池(1981)主張，國語的疑問句可分為要求回答的「徵訊問句」和不要求回答的「非徵訊問句」，修辭性問句為「非徵訊問句」的一種，即形式上雖是問句，但說話者實則表示個人的觀點或看法，語言特色為否定疑問句或者是附加問句。另一方面，Bednarek (2006, p. 67)提出六種評估英文媒體的「核心評價參數」，其中的預期性(Expectedness)的評價標記如 strikingly 和 unexpectedly，即對應到中文新聞文本中「竟（然）」、「居然」這類的評價標記，我們認為，中文新聞使用這些標記來凸顯報社的預設立場。

表 4 列出 Da San Martino *et al*. (2019)提出的 18 種宣傳手法，以及本研究發現的中文新聞特有的 2 種宣傳手法、定義及英、中文實例。我們援引 Patrona (2018)，主張帶有宣傳意味的詞語以不同形式呈現，可以是單詞、短語、子句。此外，我們將本研究對於某些手法的補充說明放入定義下方括號內以斜體字呈現。本研究參考表 4 之定義和實例進行研究語料的宣傳手法之標記工作：

表4. 英、中文新聞宣傳手法使用與實例（依Da San Martino et al., 2019 順序編排）
[Table 4. Propaganda techniques in English and Chinese new articles (In the order of that in Da San Martino et al., 2019)]

宣傳手法	定義 （*本研究的補充說明*）	英文報紙實例	中文報紙實例
1. Loaded language (LL, 評價語言)	以帶有正面／負面評價的語言描述某人、某政治群體或事件，可能是單詞、短語或子句。 （*因社論主要以評論時事為主，且具有監督執政黨的責任，通常負面語言較多。*）	"[...] a lone lawmaker's childish shouting."	負面語言的例子：「以膚淺的理由宰殺了社會正義」（聯合 2014-03-22）；正面語言的例子：「激勵愈來愈多人民抗爭」（自由 2014-03-26）。
2. Name calling or labeling (NCL, 貼標籤)	將宣傳陣營的目標貼上讀者所恐懼、厭惡、不歡迎，或者受讀者喜愛或讚揚的標籤。 （*標籤可正面或負面，以名詞片語呈現，大多是修飾結構(modifier-modified)。可以是人物的標籤，也可以是行為、政策或事件的標籤。*）	"Republican congressweasels", "Bush the Lesser."	負面標籤的例子：「馬卡茸總統」（自由 2014-03-26）;「憲政荒謬劇」（聯合 2015-08-02)；正面標籤的例子：「司法鐵漢」（聯合 2014-03-22）、「太陽花學子」（自由 2014-04-04）。
3. Repetition (RE, 重複)	反覆出現同樣的訊息使得讀者終將接受該訊息。	Farrakhan repeatedly refers to Jews as "Satan." He states to his audience [. . .] call them by their real name, 'Satan.'	本研究語料未出現此宣傳手法。
4. Exaggeration or minimization (EM, 誇大或淡化)	用過度的方法再現事件，使其看起來更好或更壞，進而影響讀者立場。	"Democrats bolted as soon as Trump's speech ended in an apparent effort to signal they can't even stomach being in the same room as the president"; "I was not fighting with her; we were	「一夕之間」（聯合 2014-06-09）。

		just playing.”	
5. Doubt (DT, 質疑)	媒體質疑政治人物或政黨可信度(credibility)，進而影響讀者立場的手法。 （*通常以問號結尾*）	A candidate says about his opponent: “Is he ready to be the Mayor?”	「哪來『深自反省』？」（自由 2015-03-20）;「難道連自己面臨什麼危機都不明白？」（聯合 2014-04-17）。
6. Appeal to fear/prejudice (ATF, 訴諸恐懼／偏見)	使用引發讀者恐懼／偏見的文字，進而影響讀者對該事件的看法。	“stop those refugees; they are terrorists.”	「統戰」（自由 2014-03-18）、「最厲的鬼魅」（聯合 2014-03-22） 。
7. Flag-waving (FW, 高舉大旗)	訴諸國民或某群體（種族、性別或政治群體）的價值偏好，以便對讀者宣揚某些想法，文字上彰顯某些（普世）價值與願景。	“stop those refugees; they are terrorists.”	「民意向背」（自由 2015-04-15）、「台灣的出路是由 ECFA 到 TPP 」（聯合 2014-03-23）。
8. Causal oversimplification (CO, 過度簡化因果)	當一項議題或事件有諸多成因，然而撰稿者卻僅咎責其一，無視其他原因，即為過度簡化因果。	“If France had not declared war on Germany, World War II would have never happened.”	本研究語料未出現此宣傳手法。
9. Slogans (SL, 呼口號)	使用簡短有力的語言來煽動觀眾情緒，其中可能帶有標籤或刻板印象。	“Make America great again!”	本研究語料未出現此宣傳手法。
10. Appeal to authority (ATA, 訴諸權威)[6]	僅僅因為某個權威或專家的支持，即認為某個主張為真。 （*引用有名的／有影響力的媒體或人物等專有名詞。*）	Monsignor Jean-Francois Lantheaume, who served as first Counsellor of the Nunciature in	「《時代》駐北京特派員指出」（自由 2014-03-26）。

[6] 本文援引 Da San Martino *et al.* (2019) 對於「訴諸權威」手法的定義：權威人士／專家／媒體／機構的主張即為真，其不同於「引用來源」之處在於，引用來源可能來自於不知名或身分模糊的人士或團體（例如台灣的媒體有時會引述網民、市民或國人等發言），而「訴諸權威」的言論則來自於權威人士／專家／媒體／機構。由於訴諸權威的重點是發言者的權威性，其發言內容的邏輯不在考量範圍內，本文因此暫不討論權威者的發言是否有「邏輯謬誤」的可能。

		Washington, confirmed that "Vigano said the truth. That's all"	
11. Black-and-white fallacy, dictatorship (BW, 非黑即白)	引導讀者認為僅存在兩種可能，但事實上仍存在其他可能。有以下兩個次類別：(1) 非 A 即 B (Black-and-white fallacy, BWF)，(2) A 或非 A (Dictatorship, DS)。	"You must be a Republican or Democrat; you are not a Democrat. Therefore, you must be a Republican"; "There is no alternative to war."	(1) 非 A 即 B：「學生的要求是畫蛇添足還是無理找碴？」（聯合 2014-03-25） (2) A 或非 A：「這不是私有化是什麼？」（聯合 2014-03-19）。
12. Thought-terminating cliché (TTC, 格言論證)	使用某些訴諸讀者常識的文字，讓讀者認為報社論點屬於常識，進而影響其立場。這些文字通常為簡短、通用的句子，為複雜的問題提供簡單的回答，讓讀者不去思考其他可能。	"it is what it is"; "you cannot judge it without experiencing it"; "it's common sense", "nothing is permanent except change", "better late than never"; "mind your own business"; "nobody's perfect"; "it doesn't matter"; "you can't change human nature."	「眾所皆知」（自由 2014-08-15）；「可謂已是在兩岸議題上的常識」（聯合 2014-03-27）。

13. Whataboutism (WH,「那又怎麼說」主義)	意指撰稿者不直接反駁對手論點，轉而攻擊對手言行不一而削弱對手立場的可信度。	Russia Today had a proclivity for whataboutism in its coverage of the 2015 Baltimore and Ferguson protests in the US, which revealed a consistent refrain: “the oppression of blacks in the US has become so unbearable that the eruption of violence was inevitable”, and that the US therefore lacks “the moral high ground to discuss human rights issues in countries like Russia and China.”	本研究語料未出現此宣傳手法。
14. Reductio ad Hitlerum (RH, 希特勒歸謬法)	暗示某些行動或想法受到大眾厭惡的團體所支持，進而說服觀眾不支持該行動或想法。	“Only one kind of person can think this way: a communist.”	本研究語料未出現此宣傳手法。
15. Red herring (RD, 原文為紅鯡魚，本文譯為「轉移焦點」)	撰稿者引入與討論議題無關的題材，以轉移讀者焦點。	“You may claim that the death penalty is an ineffective deterrent against crime – but what about the victims of crime? How do you think surviving family members feel when they see the man who murdered their son kept in prison at their expense?...”	本研究語料未出現此宣傳手法。

16. Bandwagon (BA, 從眾效應)	用「每個人都這麼做」的說詞，說服觀眾加入	"Would you vote for Clinton as president? 57% say yes."	本研究語料未出現此宣傳手法。
17. Obfuscation, intentional vagueness, confusion (OIVC, 刻意模糊與混淆)	撰稿者刻意使用不清楚的或有多重語意的文字，讓讀者有自行解讀與詮釋的空間。	"It is a good idea to listen to victims of theft. Therefore, if the victims say to have the thief shot, then you should do it."	本研究語料未出現此宣傳手法。
18. Straw man (SM, 稻草人)	用類似的說法來取代對手的主張並加以駁斥，爾後宣稱已駁倒對手的論述。	"Take it seriously, but with a large grain of salt." Which is just Allen's more nuanced way of saying: "Don't believe it."	本研究語料未出現此宣傳手法。
19. Historical allusion (HA, 引用歷史)	中文新聞獨有。使用歷史事件或人物，引發讀者對當下事件與歷史事件的類比與聯想。語言特徵為「『台（灣）版』＋歷史事件名稱」或「歷史事件名稱＋『翻版』」，或使用類比或比較文字「如」、「像」 、「相較」。	Da San Martino *et al*. (2019) 未提出此手法	「台灣版天安門事件」（自由 2014-03-26）、「仍如三十多年前美麗島事件的翻版」（自由 2014-04-04）、「有人將這場太陽花學運與一九九○年的野百合學運相較」（聯合 2014-04-05）。
20. Presupposition (PS, 預設立場)	中文新聞獨有。利用修辭性問句(rhetorical question, RQ)和評價標記(evaluative marker)來傳達報紙預設立場。其中，修辭性問句的語言特色為否定疑問句或者是附加問句，而評價標記則為中文新聞使用的「竟（然）」、「居然」這類的詞彙。	Da San Martino *et al*. (2019) 未提出此手法	(1) 修辭性問句：「豈不反諷之至？」（聯合 2014-04-29）； (2) 評價標記：「竟開放了一些不該開放的項目給中國」（自由 2014-03-18）。

本研究的標記者即為本論文作者，三位標記者皆為取得國立大學博士學位的語言學者，且專長皆涵蓋語料庫語言學和詞彙語意學。第一位標記者和資深標記者專長亦包含認知語言學，第二位標記者亦精通計算機語言學。在標記宣傳手法的過程中，三位標記者透過持續討論，確定宣傳手法定義和標記原則，並持續檢視標記結果來達到標記的一致性。

在進行標記的工作之前，我們先以中央研究院 CKIP 的斷詞系統將兩報關於太陽花學運的社論進行斷詞[7]，並以其分行結果為原則，以單行做為標記的單位。我們以【例 2】說明：

【例 2】在國會全武行的混亂中，這紙影響重大的政治與經濟開門條款，是否如國民黨團宣稱的 「視為已審查」？（自由 2014-03-18）

經中研院 CKIP 斷詞及分行的結果如下：

【例 2'】

第一行：在 國會 全武行 的 混亂 中 ，

第二行：這 紙 影響 重大 的 政治 與 經濟 開門 條款 ，

第三行：是否 如 國民 黨團 宣稱 的 「 視為 已 審查 」 ？

基於未來計算應用上的考量，我們在判斷標記時，並不考慮前後文，亦不標記跨行之宣傳手法[8]。也就是說，這三行各自形成獨立單位，例如，第一行「在 國會 全武行 的 混亂 中」，三位標記者咸認為「混亂」這個詞具備負面評價，且可單從該行判斷「混亂」為評價語言(Loaded language)，而這個詞的使用，也會影響讀者對該詞所形容事物的觀感，因此我們僅就該行中的「混亂」一詞進行標記；第二、三行亦如此。

再者，一行內部可能出現多於一個的宣傳手法標記，也可能包含多於一種的宣傳手法，例如，【例 3】包含了兩個評價語言的標記，即「膚淺」和「宰殺」；【例 4】則出現了一個貼標籤的標記「九趴總統」，和兩個評價語言標記「執迷不誤」與「一意孤行」：

【例 3】以<u>膚淺</u>的理由<u>宰殺</u>了社會正義（聯合 2014-03-22）

【例 4】<u>九趴總統</u>至今仍<u>執迷不誤</u>*[sic]*、<u>一意孤行</u> 。（自由時報 2014-04-04）

標記程序分成三階段，在第一階段時，第一和第二標記者根據前述表 4 之宣傳手法定義對同一份文件分別進行標記。第二階段中，兩位標記者與資深標記者討論，並請資深標記者針對不一致處做最終版本之標記。三位標記者決議，標記者間一致性以最嚴謹的三位標記者的宣傳手法標記皆一致的百分比(unanimous percentage)進行評估。

---

[7] 為便利單詞的計算、標記與量化，本文在分析新聞語料前仍須先以斷詞系統處理語料。

[8] 本文做為中文政治宣傳手法標記之先導研究，現階段依循英文之類似研究，僅以單一行內之前後文脈絡，做為判斷宣傳手法標記之依據。若考慮文章段落層次之前後文脈絡，人工標記將需要更多時間，計算任務上亦須先修改引用的 Da San Martino *et al*. (2019)之模型，以應用於訓練根據段落內含跨行線索之複雜層次標記，否則將無法凸顯長脈絡線索標記之優勢。後續的研究中，我們將會考量長脈絡運用於宣傳手法標記之可能性。

為了加快標記速度與增加標記規模，初步人工標記完前 10%的文章後，以 Bootstrap 方式將人工標記做為種子，比對其他 90% 的文章是否含有與人工標記相同之內容，若分行中出現曾標記過之單詞／短語／子句，即將同一單詞／短語／子句視為相同宣傳手法，做為第三階段之標記資料。例如：在第二階段三位標記者一致認為「癱瘓」在【例 5】中為評價語言：

【例 5】抗爭運動不僅癱瘓國會運作（聯合 2014-03-30）

Bootstrap 即以「癱瘓」一詞，比對 90%未經人工標記的文章，並將【例 6】－【例 12】行段中的「癱瘓」標記為評價語言：

【例 6】向來是癱瘓的時間比運作的時候多；（聯合 2014-04-09）

【例 7】皆有假藉抗爭的道德性來癱瘓體制的危險。（聯合 2015-03-18）

【例 8】並幾乎癱瘓了馬政府的執政能力；（聯合 2015-08-03）

【例 9】學生占領國會、癱瘓政府的行動 ，（聯合 2015-10-22）

【例 10】因為這將癱瘓政府機構的正常運作。（聯合 2016-05-26）

【例 11】當時學生以占領議場癱瘓國會，（聯合 2016-07-01）

【例 12】為何當時學生癱瘓國會被捧為英雄，（聯合 2016-11-04）

經過 Bootstrap 擴展的標記資料，以人工檢核方式提高標記的品質。在此過程中，我們也發現以單詞為主的標記手法 Bootstrap 的成效較佳，例如貼標籤、誇大或淡化、質疑、訴諸恐懼、高舉大旗、引用歷史…等，而評價語言和預設立場中，部分短語（如「為反對而反對」）或子句（如「你們心中真的還有別人嗎？」）則因斷為不同單詞，需要在人工檢核時修正，這可列為未來機器學習訓練特徵的考量之一。

為確保標記資料集的效能，我們進行先導實驗以為外部評估。我們選取第一階段經由人工標記、包括含有宣傳手法與不含宣傳手法的所有行段進行機器學習，並分別在這兩批資料中，選取含有宣傳手法標記與不含宣傳手法標記的行段之各 70%，作為機器訓練的語料，透過單純貝式分類器(Naïve Bayes classifier)與詞袋特徵(bag of words feature)進行機器學習。在機器學習完成後，先清除這兩批資料另外的 30%行段中的所有標記，並讓機器判讀一行段是否包含宣傳手法[9]。實驗結果將呈現於第 4 章「結果與討論」。

4. 結果與討論 (Results and Discussion)

本研究的語料經中研院 CKIP 斷詞系統處理後，得到自由時報社論共 17,098 行，聯合報社論共 17,888 行。標記後發現，自由時報共有 2,312 行含有宣傳手法標記，聯合報共有 2,413 行含有宣傳手法標記。由於本標記工作較傳統分類問題複雜，第一階段在自由時報前 8 篇共 1,073 行中，第一位標記者共標記了 445 個宣傳手法，第二位標記者則有 86 個

[9] 基於時間考量，現階段的先導實驗，我們先以機器判讀一行段中是否存在宣傳手法為原則。機器判讀宣傳手法種類之研究，需要更多時間，我們留待下一階段的研究執行。

宣傳手法標記，其中有 594 行兩位標記者皆認為不含宣傳手法，並有 28 個兩位標記者標記一致的宣傳手法；聯合報前 15 篇共 1,825 行中，第一位標記者標出了 577 個宣傳手法，第二位標記者有 104 個宣傳標記。第一階段兩位標記者間的一致性(inter-annotator agreement)僅達 63.3%[10]。三個階段分別處理的標記數整理如表 5。

表 5. 三個階段分別處理的標記數量
[Table 5. Number of markers at the three stages]

	自由時報	第一位標記數	第二位標記數	聯合報	第一位標記數	第二位標記數
第一階段	1,073	449	86	1,825	577	104
第二階段	同第一階段	-	-	同第一階段	-	-
第三階段	2,312	-	-	2,413	-	-

關於第一階段兩位標記者宣傳手法標記數量上的差異，來自於第一位標記者認為單詞／短語／子句皆可傳達某種宣傳手法，而第二位標記者僅針對單詞進行標記。第二階段兩位標記者與資深標記者討論，對於帶有宣傳手法的詞語進行定義（即單詞／短語／子句皆可傳達某種宣傳手法），並請資深標記者針對不一致處做最終版本之標記，最後決議標記者間一致性以最嚴謹的三位標記者的宣傳手法標記皆一致的百分比(unanimous percentage)進行評估，則分別為自由時報的 14.9%（共有 67 個三位標記者皆一致的宣傳手法標記）以及聯合報的 9.9% (57 個完全一致的宣傳手法標記)。最後再進行第三階段以 Bootstrap 方式擴增資料標記規模，並以人工校正[11]確保 Bootstrap 標記品質。

人工校正的筆數如表 6 所示，以 Bootstrap 方式擴展但遺漏處，再以人工新增筆數，分別為自由時報 266 筆以及聯合報 328 筆；另自由時報有 77 筆、聯合報有 105 筆 Bootstrap 不盡正確處再以人工更正，其中含 Bootstrap 誤標之 59 筆（自由時報）與 66 筆（聯合報）。第一階段的標記最後成為最終版本之百分比達 85.2%（自由時報）[12]以及 82.1%（聯合報）。

---

[10] 本研究以第一標記者為基準，與第二標記者對同一行標記結果是否一致的百分比。本研究使用的新聞文本大部分的行數並未使用宣傳手法，我們僅針對含有宣傳手法的行段進行統計分析。

[11] 使用 Bootstrap 擴大後，由於 Bootstrap 可能會有缺漏之處，因此需要以人工校正來檢視，人工校正階段在經由三位標記者同意後確認。

[12] 由於本文的 Bootstrap 方式純粹根據第二階段的人工標記最終版本，僅比對相同單詞／短語／子句即進行同樣的標記。原第二階段的人工標記（視為標準答案，100%正確），加上 Bootstrap 後正確的總筆數為 1,969 筆，其他修正的筆數則為 266+77=343 筆，因此憑藉第二階段標記（含原人工標記數）而產生的標記正確率為 1,969 / (343+1,969) = 85.2 %。此百分比亦與 Bootstrap 方法的表現相關。

表 6. 人工校正兩報社論之宣傳手法標記之統計
[Table 6.The statistic of the manual correction of the propaganda technique markers in the Liberty Times and the United Daily News]

校正方式	自由時報	聯合報
新增(A)	266 (11.5%)	328 (13.2%)
更正(Y)	77 (3.3%)	105 (4.4%)
正確	1,969 (85.2%)	1,980 (82.1%)
總計	2,312 (100.0%)	2,413 (100.0%)

表 7. 自由時報社論各宣傳手法標記之 ***Bootstrap*** *正確率統計（依照總筆數排序）*
[Table 7. The statistic of the Bootstrapped markers of the propaganda techniques of the Liberty Times (In the order of the total number of each technique)]

	宣傳手法	更正筆數	新增筆數	正確筆數	總筆數	正確率
1	評價語言	15	199	1,116	1,330	83.9%
2	高舉大旗	0	16	416	432	96.3%
3	貼標籤	2	39	90	131	68.7%
4	非黑即白	0	3	127	130	97.7%
5	訴諸恐懼	0	4	123	127	96.9%
6	預設立場	0	4	42	46	91.3%
7	訴諸權威	1	0	34	35	97.1%
8	引用歷史	0	0	12	12	100.0%
9	格言論證	0	0	6	6	100.0%
10	質疑	0	0	3	3	100.0%
11	誇大或淡化	0	1	0	1	0.0%
12	其他(X)	59	0	0	59	•
	總和	77	266	1,969	2,312	85.2%

表 7、表 8 分別呈現兩報各項宣傳手法之 Bootstrap 正確率（與人工檢核比較），其中更正筆數表示 Bootstrap 標記錯誤、經人工更正成正確的標記筆數；新增筆數為 Bootstrap 未比對出之標記，經人工新增之標記筆數；其他未經變動的筆數皆為正確筆數，總筆數則代表最終標記版本中，該宣傳手法出現的筆數，正確率之算法則為正確筆數除以總筆數。表格中第 12 類標記為其他(X)者，表示以 Bootstrap 方式標記錯誤，該行並未含有任何宣傳手法。

以表 7 自由時報社論中評價語言宣傳手法的標記為例，有 15 筆被 Bootstrap 方式

錯標為其他手法，後來經人工更正為評價語言標記；另 Bootstrap 遺漏 199 筆評價語言標記，而在最後人工檢核階段，經人工新增為評價語言標記；剩下 1,116 筆為以 Bootstrap 方式比對後完全正確，未再經人工變動的正確筆數；最後自由時報社論中，總共標記了 1,330 筆評價語言，正確率為 1,116 / 1,330 = 83.9%。

表 8. 聯合報社論各宣傳手法標記之 Bootstrap 正確率統計（依照總筆數排序）
[Table 8. The statistic of the Bootstrapped markers of the propaganda techniques of the United Daily News (In the order of the total number of each technique)]

宣傳手法	更正筆數	新增筆數	正確筆數	總筆數	正確率
1 評價語言	13	278	1,449	1,740	83.3%
2 非黑即白	15	0	257	272	94.5%
3 預設立場	0	9	146	155	94.2%
4 貼標籤	10	23	26	59	44.1%
5 訴諸恐懼	0	4	31	35	88.6%
6 高舉大旗	1	3	27	31	87.1%
7 誇大或淡化	0	0	17	17	100.0%
8 格言論證	0	4	9	13	69.2%
9 引用歷史	0	1	10	11	90.9%
10 質疑	0	3	7	10	70.0%
11 訴諸權威	0	3	1	4	25.0%
12 其他(X)	66	0	0	66	
總和	105	328	1,980	2,413	82.1%

從表 7 和表 8 中可見，兩報社論以 Bootstrap 比對方式標記的平均正確率皆達八成以上（自由時報為 85.2%，聯合報為 82.1%），惟其他未經 Bootstrap 的行數尚有可能含有遺漏之標記。

我們發現常見的宣傳手法之比例如表 9 所述。相較於 Da San Martino *et al.* (2019, p. 5641) 在英文新聞的標記結果，中文報紙和英文報紙的「評價語言」的手法出現頻率皆屬最高，然而其餘手法卻呈現不同的面貌：排除「其他」這個類別，英文報紙第二、第三和第四高比例的手法分別為「貼標籤」、「誇大或淡化」和「質疑」，而中文報紙第二、第三和第四高比例的手法則是「高舉大旗」和「非黑即白」，以及未在英文新聞中

出現的「預設立場」。

表 9. 中、英文新聞之常見宣傳手法與比較（依照中文的百分比排序）
[Table 9. The comparison between the propaganda techniques employed in the Chinese and English news articles (In the order of the percentage of each technique in the Chinese news articles)]

宣傳手法	自由時報	聯合報	總標記數	百分比	Da San Martino et al.	百分比
1. 情緒語言	1,330	1,740	3,070	65.0%	2,547	34.1%
2. 高舉大旗	432	31	463	9.8%	330	4.4%
3. 非黑即白	130	272	402	8.5%	134	1.8%
4. 預設立場	46	155	201	4.3%	-	-
5. 貼標籤	131	59	190	4.0%	1,294	17.3%
6. 訴諸恐懼	127	35	162	3.4%	367	4.9%
7. 訴諸權威	35	4	39	0.8%	169	2.3%
8. 引用歷史	12	11	23	0.5%	-	-
9. 格言論證	6	13	19	0.4%	95	1.3%
10. 誇大或淡化	1	17	18	0.4%	571	7.6%
11. 質疑	3	10	13	0.3%	562	7.5%
12. 其他（X）	59	66	125	2.6%	1,411	18.9%
總數	2,312	2,413	4,725	100.0%	7,480	100.0%

外部評估的先導實驗結果顯示，以單純貝式分類器搭配詞袋特徵進行訓練後，機器判讀的準確率可達 74.26%，顯示本宣傳手法標記資料確實有助於機器自動判讀宣傳手法。

5. 結論 (Concluding Remarks)

本文已回答了前述三個研究問題，首先我們針對中文的政治新聞提出 11 種宣傳手法，其中「引用歷史」和「預設立場」有別於英文，為中文新聞獨有之宣傳手法。而根據標記自由時報與聯合報共 228 篇(360,999 詞)太陽花學運相關社論的結果統計，中英文新聞除了皆最常使用「評價語言」手法之外，其他宣傳手法呈現完全不同的分佈。惟本研究之中文新聞取材僅限於社論此一文類，與 Da San Martino *et al*. (2019)所採用宣傳型新聞網站之新聞文類不盡相同，因而此處的中、英統計數據之比較會有所限制。未來在文類選擇上，我們將選取更高可比性的文類，並將新聞主題變項、報導發布時間與文章篇數等條件納入選材之考量。

以 Bootstrap 比對方式擴增標記之正確率兩報皆達八成左右，但尚有其他未經 Bootstrap 比對出之行數，也可能使用宣傳手法。我們未來將繼續檢視本宣傳手法標記資料中，是否尚有其他未比對出之標記，並進行評估。

本文做為首篇標記中文宣傳手法之研究，目前僅初步評測標記者間針對同一行的標記是否一致的百分比，尚未考慮其他統計學者所提出之 γ 評測，針對不同標記者部分標記重疊進行評測。此為本初步研究之限制，後續研究在本標記手法定義之基礎上，可專注於標記者事前的訓練，並採取單一階段之標記流程，以達較高的標記品質。再者，由於宣傳手法標記之複雜度，本文除了基本的標記一致性之百分比，僅將人工標記之宣傳手法資料用 Bootstrap 方式擴展資料集，最後進行人工檢核。關於運用機器計算模型於此類型標記之可能性，將列為未來繼續研究之議題。

本文人工標記之宣傳手法語料，已開放用於偵測宣傳手法以及相關研究[13]，亦可用於媒體中立程度之評估。本研究未來將延續目前的宣傳資料標記，並擷取其中常見之語言特徵（如關鍵詞頻或句法結構），自動判讀其他未標記新聞，並偵測其所採用之宣傳手法。

致謝 (Acknowledgements)

本論文承蒙國立政治大學資訊科學系江玥慧老師協助分析英文新聞之宣傳手法、科技部計畫 MOST 109-2811-H-004-503、109-2410-H-004-163- 與 110-2811-H-007-502 補助，以及三位匿名審查委員的指導，特此致謝。

參考文獻 (References)

Arefi, M. N., Pandi, R., Crandall, J. R., Tschantz, M. C., Fu, K.-W., Shi, D. Q., & Sha, M. (2019). Assessing post deletion in Sina Weibo: Multi-modal classification of hot topics. In *Proceedings of the 2nd Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, 1-9. https //doi.org/10.18653/v1/D19-5001

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555-596. https://doi.org/10.1162/coli.07-034-R2

Barrón-Cedeño, A., Jaradat, I., Da San Martino, G. & Nakov, P. (2019). Proppy: Organizing the news based on their propagandistic content. *Information Processing and Management, 56*(5), 1849-1864. https://doi.org/10.1016/j.ipm.2019.03.005

Bednarek, M. (2006). *Evaluation in Media Discourse: Analysis of a Newspaper Corpus.* New York: Continuum.

Bensa, C. P. & Wijaya L. (2017). Media propaganda techniques in the South China Sea dispute. *Jurnal Komunikasi ISKI*, *02*(01), 1-5. https://doi.org/10.25008/jkiski.v2i1.86

Bussemer, T. (2005). *Propaganda: Konzepte und Theorien*. Wiesbaden: VS Verlag für Sozialwissenschaften.

[13] 本標記資料集可於此網址下載 http://simonshih.tw/propaganda/propaganda_phrase_list.{LT,UDN}.bootstrapped.tsv。

Da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R., & Nakov, P. (2020). SemEval-2020 Task 11: Detection of propaganda techniques in news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, 1377-1414.

Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., & Nakov, P. (2019). Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 5636-5646. https://doi.org/10.18653/v1/D19-1565

Diggs-Brown, B. (2011). *Strategic Public Relations: An Audience-Focused Approach.* Boston: Wadsworth Cengage.

Fowler, R. (1991). *Language in the News: Discourse and Ideology in the Press.* Routledge.

Fairclough, N. (1995). *Critical Discourse Analysis.* Longman.

Herman, E.S. & Peterson, D. (2010). *The Politics of Genocide*. New York, NY: Monthly Review Press.

Hodge, B. (1979). Newspapers and communities. In Fowlder, R., Hodge, B., Kress, G., & Trew, T. (eds.) *Language and Control,* (pp.157-174). Routledge & Kegan Paul.

Horne, B. D., Dron, W., Khedr, S., & Adali, S. (2018). Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, 518-527.

Keeble, R. (1997). *Secret State, Silent Press: New Militarism, the Gulf and the Modern Image of Warfare*. Luton: John Libbey Media.

Knockel, J., Crete-Nishihata, M., Ng, J. Q., Senft, A., & Crandall, J. R. (2015). Every rose has its thorn: Censorship and surveillance on social video platforms in China. In *Proceedings of the 5th USENIX Workshop on Free and Open Communications on the Internet*, 1-10.

Kromrey, H. (2009). *Empirische Sozialforschung, 12th edn*. Stuttgart: Lucius & Lucius.

Lasswell, H. D. (1927). *Propaganda Technique in the World War.* Michigan: Peter Smith.

Lee, A. M., & Lee, E. B. (1939). *The Fine Art of Propaganda*. New York: The Institute for Propaganda Analysis.

Lin, Y., & Hsieh, J. F. (2017). Change and continuity in Taiwan’s public opinion on the cross-Strait economic interactions. *Journal of Asian and African Studies, 52*(8), 1103-1116. https://doi.org/10.1177/0021909616649210

Mathet, Y., Widlocher, A., & Métivier, J.-P. (2015). The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics, 41*(3), 437-479. https://doi.org/10.1162/COLI_a_00227

Mohammad, S. M., Sobhani, P., & Kiritchenko, S. (2017). Stance and sentiment in tweets. *ACM Transactions on Internet Technology, 17*(3), 1-23. https://doi.org/10.1145/3003433

Patrona, M. (2018). Mediated conversation and political propaganda: A discourse analysis of news practices on Greek prime-time news. *European Journal of Communication*, *33*(2), 185-199. https://doi.org/10.1177/0267323118760321

Riegel, O. W. (1935). Propaganda and the press. *The ANNALS of the American Academy of Political and Social Science, 179,* 201-210.

Scott, W. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*(3), 321-325.

Smith, B. L. (2020). Propaganda. In *Britannica*. Retrieved from www.britannica.com/topic/propaganda on 18 February 2021.

van Dijk, T. (1988). *Discourse as News*. Lawrence Erlbaum Associates, Inc.

van Dijk, T. (1995). Opinions and ideologies in editorials. Paper presented at *the 4th International Symposium of Critical Discourse Analysis*, Athens, Greece.

van Leeuwen, T. (2008). *Discourse and Practice. New Tools for Critical Discourse Analysis.* Oxford University Press.

Wang, W. (2007). *Genre across Languages and Cultures: Newspaper Commentaries in China and Australia*. VDM Verlag Dr. Muller.

Weston, A. (2018). *A Rulebook for Arguments, 5th edn*. Cambridge: Hackett Publishing.

Zollmann, F. (2017). *Media, Propaganda and the Politics of Intervention*. New York, NY: Peter Lang.

Zollmann, F. (2019). Bringing propaganda back into news media studies. *Critical Sociology*, *45*(3), 329-345. https://doi.org/10.1177/0896920517731134

朱灼文（2003）。**社論的論證結構分析**。國立政治大學新聞研究所碩士論文。[Chu, C.-w. (2003). *The Analysis of Argument Structures in Editorials*. M.A. Thesis, National Chengchi University.]

洪見合（2014）。**中英文新聞社論之跨文化篇章分析：系統功能語言學觀點**。國立高雄應用科技大學應用外語系暨英語專業溝通與教學科技碩士論文。[Hung, C.-H. (2014). *A Genre Analysis of Chinese and English Newspaper Editorials across Cultural Variations: A Systemic Funcational Perspective*. M.A. Thesis, National Kaohsiung University of Science and Technology.]

游承季（2001）。**大陸經濟改革的內容和宣傳策略研究(1992-1996)**－以「人民日報」社論內容為例。國立政治大學新聞研究所碩士論文。[You, C.-C. (2001). *On the Contents and Propaganda Strategies of China's Economy Reformation (1992-1996) in the Editorials of the People's Daily*. M.A. Thesis, National Chengchi University.

湯廷池（1981）。國語疑問句的研究。**師大學報** 26, 1-59。[Tang, T.-c. (1981). On the interrogative sentences in Chinese. *Journal of National Taiwan Normal University*, 26, 1-59.]

賴晉（1965）。**人民日報社論的反美宣傳分析**。國立政治大學新聞研究所碩士論文。[Lai, M. (1965). *Analyzing the Anti-USA Propaganda in the Editorials of the People's Daily*. M.A. Thesis, National Chengchi University.]

附錄1 (Appendix 1)

第一階段人工標記的自由時報宣傳手法以及相對應的單詞／短語／子句（以英文字母排序）

Appeal to authority (ATA, 訴諸權威)：《 彭博 商業 周刊 》《 時代 》「 紐約時報 」 紐約時報《 紐約時報 》中國 移居 美國 的 經濟 社會 學者 華爾街日報「 經濟 學人 」 希拉蕊 傳統 基金會 華爾街日報 國際 媒體 威爾遜 華爾街日報 台灣 公報 希拉蕊 ・ 柯林頓 鄭秀玲 經濟 學人 雜誌 華爾街日報 長期 關心 台灣 的 專家 甄妮

Appeal to fear/prejudice (ATF, 訴諸恐懼)：一 中

Black-and-white fallacy, dictatorship (BW, 非黑即白)：任何 一 個 上軌道 的 國家 絕不會 允許 在 毫無 討論 、 不准 修正 改良 的 粗暴 過程 中 絕不 絕非 必須

Doubt (DT, 質疑)：其 此 之 謂 乎 ？ 哪來「 深 自 反省 」？ 難道 是 這樣 的 想法 嗎 ？

Flag-waving (FW, 高舉大旗)：民主 價值 民意 心聲 公民 覺醒 民意 國家 安全 公眾支持 捍衛 民主 普遍 心聲 民意 民調 正常 民主 程序 國家 安全 民意 基礎 主權國安

Historical allusion (HA, 引用歷史)：美麗島 事件 天安門 事件 「 引 清兵 入關 」

Loaded language (LL, 評價語言)：血腥 毀滅 不 倫 「 氣 爆扯 服貿 」 食言而肥 紙上談兵 一意孤行 ， 吐 槽 任 人 宰割 馬王 惡鬥 內鬨 神明 閉門思過 熱潮 幹譙向 中國 傾斜 卸責 搪塞 令 人 感嘆 可悲 食髓知味 食髓知味 ， 敲詐 併吞 門戶大開 罔顧 矛盾 玩弄 慢半拍 一意孤行 執迷 不誤 寶貴 的 誘使 陷阱 震撼 低迷罔顧 濫行 賤 招 撐腰 低聲下氣 無能 來 洗 人民 的 腦 臭 酸 的 暴力 背對 人民目 無 頭家 戲本 粗暴 強 渡 關山 操控 傷害 優哉游哉 ， 無能 無心 偷偷 覺醒 傷害 惡評 爽 歪歪 障礙 公民 覺醒 遍 地 開花 威脅 衝擊 危機 專制 權謀 算計 剝奪 「 無能 」 助成 失靈 嫁禍 製造 事端 報復 粗糙 輕忽 暴力 自恃 自我 吞噬 賠上 黑箱 語言 暴力 墮落 垂涎三尺 變 不 出新 把戲 執迷不悟 五花大綁 嘈雜 紛亂詬病 封殺 石破天驚 壓制 動盪 嚴重 淪為 逆來順受 撕毀 崩盤 譴責 嚴厲 惡評 混亂 有志一同 大放厥詞 。 公民 覺醒 熱力 啟發 壯志 宰殺 不可 原諒 險境 ， 盜取破產 當頭棒喝 傷害 嚴峻 復健 永無寧日 癱瘓 強推 賣國 叛國 暴力 霸凌 分官 分贓 搶 立委 情何以堪 。 黑 箱 危機 自 失 受 威脅 風險 困境 抹黑 爛 講義 抹黑謊言 羅織 罪名 嚴厲 譴責 不實 反擊 揭露 破解 過氣 放任 媚共 恥 卑躬屈膝 暗潮 浪花 傷害 反感 言行 不一 行動 暴力 中肯 一意孤行 圖謀 污名 化 杯葛 唯命是從 對付 打壓 併吞 蕩然無存 毫不 在意 。 奮起 漁獵 激勵 黑箱 蹂躪 朝氣 清新愚蠢 致命 錯誤 恐嚇 毀壞 伎倆 黑道 恐嚇 無比 威力 偷渡 濫 可笑 硬 愚蠢 粗暴粗糙 蠻橫 「 改變成 真 」 惡果 憂慮 封殺 回 嗆 投票 部隊 招數 玩弄 虛假 哈腰諂媚 「 以 商 逼 政 」 虛晃 一 招 罔顧 民怨 罔顧 恐嚇 牌 施暴 開門揖盜 肆 背離 遭 打 臉 威脅 唬弄 咎由自取 轉捩點 慷慨 陳詞 見賢思齊 。 諷刺 謊話 連篇

肖 話 顏面 盡失 灰頭土臉 輸掉 一廂情願 蠻幹 軟土 深 掘 跳針式 喊話 傲慢 糞土 危機 國民黨 「 暴政 」 怵目驚心 落伍 向 中國 靠攏 圍 事 失敗 的 最 爛 開門揖盜 恐怖 悲哀 沒 救 的 高高在上 逢迎 一針見血 惡果 煙霧 門戶 洞開 ， 害台 權力 的 傲慢 束手無策 目 無 災民 離題 挺身而出 無能 不當 專制 牢籠 衝突 冥頑不靈 交叉 感染 遺毒 變 不 出新 把戲 偷渡 門戶 洞開 蠶食 鯨吞 「 獨裁 」 被迫 「 無能 」 強推 黑 箱 沾沾自喜 日薄西山 獨沽 中國 一味 騙局

Name calling or labeling (NCL, 貼標籤)：「 半 分 忠 」 民怨 火山 「 鐵蹄 」 「 太陽餅 馬戲團 」 「 馬卡茸 總統 」 「 現代 吳鳳 」 勇敢 婦人 中國 台商 投票 部隊「 抓漏 大隊 」元凶 九 趴 總統 即將 下台 的馬英九 們 馬英九 們 馬 黨國 無心 也 無能 的 馬 政府 傾中 政策 老共 追求「 終極 統一 」的 馬英九「 今日 香港 」「 一 條 龍 」 九 趴 總統 國會 全武行 向 中國 傾斜 政策 明日 台灣 」 幫腔 幫閑 者「 台灣版 天安門 事件 」未來 的 主子 鳥籠 公投「 公民 不 服從 運動 」背離 民意 的 政府 黑 箱 服 貿 ， 把 國旗 收 起來 的 小 官 ， 不 適格 不適任 者 黨國 輿論 製造術 反 民主 鬧劇 有 勝算 者 有識之士 列寧式 黨 機器 馬黨國 「 天龍國 」 黨 外 精英 「 中國 軍閥 鐵蹄 」 醜 角色 紅頂 商人 元凶 幫凶 共匪 太陽 花 學子 老把戲「 婉君 」九 趴 總統 長期 關心 台灣 的 專家「 暴鎮 」 中國黨 「 無能 領導 者 」 違法 總統 這 個 「 自絕 於 國人 」 的 政府 雙手 獨攬 權力 的 馬英九 「 暴鎮 」 馬 黨國 老狗 馬英九 們 ， 黑 箱 服 貿

Presupposition (PS, 預設立場)：竟然 不 是 更 應該 閉門思過 、 內慚 神明 嗎 ？ 台灣 經濟 會 愈 拚 愈 倒退 嗎 ？ 竟 實質 薪資 所得 會 愈 拚 愈 回去 嗎 ？ 竟是否 如 國民 黨團 宣稱 的 「 視為 已 審查 」 ？ 武官 還 怕 甚麼 ？ 滿意度 會一路 滑到 九 趴 嗎 ？ 竟 誰 會 看 不 懂 、 想 不 清楚 ？ 貧富 差距 會 持續擴大 嗎 ？ 這 叫 甚麼 ？ 這 豈不 說明 M 503 航路 對 我國 飛 安 與 國安 確實存在 威脅 ？ 還 有 多少 看 不 到 卻 更 恐怖 的 事情 正在 發生 ？ 難道 這 是人民 的 「 暴力 杯葛 」 嗎 ？ 馬英九 不 也 是 如此 心態 嗎 ？

Thought-terminating cliché (TTC, 格言論證)：不 誠 無物 眾所周知 ， 答案 當然 是 否定 的 自不待言

附錄2 (Appendix 2)

第一階段人工標記的聯合報宣傳手法以及相對應的單詞／短語／子句（以英文字母排序）

Appeal to authority (ATA, 訴諸權威)：嚴 長壽 彭淮南 施振榮 童子 賢 附庸 化 「 害台 禍台 」 邊緣化 絕路 民主 倒退 邊緣化 最 厲 的 鬼魅

Black-and-white fallacy, dictatorship (BW, 非黑即白)：沒有 任何 國家 能 容許 完全 是朝野 立委 爾虞我詐 、 機關 算 盡 的 結果 。不可能 靠 著 街頭 抗議 或 國際 鎂光燈 解決 絕不 這 不 是 民進黨 的 「 私有化 」 是 什麼 ？ 這 個 國家 豈不 注定 要 在 想像 與 現實 的 兩極 間 自我 拉扯 ？ 絕不 學生 的 要求 是 畫蛇添足

還是 無理 找碴 ？ 抗中 或 害台 ？ 絕不 這 不 是 民進黨 的 「 私有化 」 是 什麼 ？ 必須

Doubt (DT, 質疑)：卻 仍 要 將 台灣 帶上 絕路 ？ 卻 連 這 點 想法 也 容不下 ？ 又 在 維護 什麼 樣 的 司法 正義 ？ ？ 又 如何 贏得 民眾 認同 ？ 會不會 又 要 商請 俄國 寬限 ， 為 台灣 爭取 利益 ？ 烏克蘭 財政 會不會 崩潰 ， 在 彰顯 什麼 樣 的 法律 價值 ？

Exaggeration or minimization (EM, 誇大或淡化)：轉瞬 一 夕之間 永遠

Flag-waving (FW, 高舉大旗)：「 媽媽 反核 能 」 「 學生 反 服貿 」 台灣 的 出路 是 由 ＥＣＦＡ 到 ＴＰＰ 。 國會 正 是 民主 憲政 中 由 選舉 產生 的 最 高 民意 機關 ， 國會 象徵 的 就是 民意 。 「 人民 」 「 捍衛 民主 」 民主 大 教室 呵護 我們 得來 不易 的 自由 「 台灣 全球化 機遇 」 反映民意 承載民意 讓 學生們 平安 地 回到 學校 及 家庭 吧 ！ 「 全民 總統 」 民主 精神 「 人民 」

Historical allusion (HA, 引用歷史)：「 茉莉花 革命 」 「 五二○ 農運 」 被 釘上 十字架 泰國 的 紅 衫軍 和 黃 衫軍 之 爭 烏克蘭 十 年前 的 橘色 革命 野百合

Loaded language (LL, 評價語言)：無可 逆轉 伎倆 權謀 鬥爭 不攻自破 底蘊 崩盤 戲碼 虛矯 「 政府 的 失 能 」 濫用 磨刀 霍霍 蒙上 陰影 上演 攻防爛戲 抄襲 搬弄 抹 除 牛頭不對馬嘴 逾越 徒 添 困惑 矯情 斥為 難堪 狠狠 長才 趨吉避凶 排擠 悲劇 誤解 畫蛇添足 危害 風暴 惶恐 焦慮 狂暴 步步 退讓 淪陷 羞愧 撕裂 窄巷 宰殺 膚淺 以 製造 騷亂 自得 ， 鬧劇 姑息 漠然 潰敗 仇恨 焦慮 、 猜疑 揭竿 而起 崩毀 矛盾 被 踐踏 短短 絕無 可能 喪志 喪膽 戲碼 杯葛 脫序 衝突 暴力 過激 惡化 凶險 憂慮 脫序 汗顏 錯覺 今非昔比 撇開 故意 漫天 喊價 扭曲 戰場 一搭一唱 壟斷 嚴重 衝擊 區區 妨礙 嚴重 脫軌 害台 禍台 傷痛 絕路 以 占領 或 攻擊 政府 機構 為 樂 動輒 投鼠忌器 氣焰 高漲 頤指氣使 害台 禍台 輕輕 撇 到 一 旁 不理 口惠 落空 喪 權 失土 大 毒瘤 毒瘤 伎倆 唾 餘 權謀 鬥爭 陷 奪取 託詞 反智 井然有序 摧毀 大 爆發 醜化 藉口 亂刀 命 喪 「 民主 的 失控 」 得寸進尺 崩裂 蹉跎 結黨營私 壟斷 逼 得 得寸進尺 光榮 侵犯 光榮 熱鬧 無理 找碴 畫蛇添足 義憤 難安 玩弄 於 股掌 之上 權謀 鬥爭 耽溺 見獵心喜 機關 算 盡 爾虞我詐 本末倒置 錯愕 驚駭 不成體統 正大 鬥垮 無效 自傷 動輒 撕裂 傷害 殘害 聲望 連挫 干擾 無底洞 失格 羞辱 國家 、 羞辱 民主 蕩然無存 兵疲 馬 困 怒火 引爆 勃然大怒 霸凌 鬥爭 絕路 不堪 想像 得不償失 。 得寸進尺 不見天日 壟斷 狡猾 失之 偏頗 狗尾續貂 大話 為 反對 而 反對 走投無路 流失 硬生生 咬死 害台 抗中 癱瘓 拖 了 指使 破壞 憲政 挾制 民意 伏擊 杯葛 鬥爭 戰場「 不 作為 」打發 癥結 瘖啞 難辭其咎 一廂情願 強推 政策 悍然 危機 一針見血 混亂 程序 杯葛 令 人 痛心 的 嚴重 政治 災難 狹隘 拉高 身段 難堪 破裂 ， 根本 不忍卒睹 醜態 根本 剝奪 集體 暴力 根本 坐收漁利 下 了 如此 重 手 交易 拖延 杯葛 前車之鑑 光榮 煽風 點火 為 鬥爭 而 鬥爭 視 人民 為 政黨 可 驅遣 之 「 工具 」 「 霸權 」 大剌剌

併吞 令 人 三 嘆 抹煞 賠上 停損點 唾面自乾 地 耗 著 惡性 循環 陷 流產 政治 危機 爆發 杯葛 鎮日 霸占 倉皇 利益 交換 密室 協商 煽動 脫序 唱和 糾纏 軟弱 軟弱 可 欺 暴力 蒙上 了 譴責 倒退 不能 動彈 被 卡住「 反 民主 」「 黑箱 」，錯誤 百出 諷刺 脫序 袖手旁觀 憂心忡忡 嘆為觀止 予取予求 窘境 內亂 困窘 迫使 讓 人 遺憾 亂子 嚴酷 挾制 自我 拉扯 漠視 風暴 混亂 衰變 通病 恥 悲劇 迅雷不及掩耳 地 荒謬 政府 的 失 能 決裂 扭曲 拘泥 不 化 動輒 一黨之私 汲汲 罔顧 倒錯 逼 得 黨政 不 分 違論 害台 禍台 幻影 傷痕 卸下 踐踏 肆無忌憚 隨便 「 干涉 步步 退讓 狂暴 反對 置若罔聞 受挫 步步 退讓 一敗塗地 賠上 薄弱 羞辱 驅逐 衝撞 葬送

Name calling or labeling (NCL, 貼標籤)：「 司法 鐵漢 「 零 容忍 」 的 社會 「 包養 條約 」， 正義 之 師 揭弊 者 關說 者「 台灣 社會 良心 」的 意見 領袖 意見 領袖 鬥獸場 「 解嚴 後 首 波 民主 世代 」 馬戲團 鬥獸場 「 以 商 圍 政 」之 術 霸占 國會 者 野百合 北極 狼 忠誠 堪虞的 人 民主 大 教室 關說 者 活 在 「 象牙塔 裡 」 的 人 ？ 天王 們 法匠 結黨 結派 的 政客 服貿 黑箱 「 反中 抗中 」 的 風暴 棉花糖 最 厲 的 鬼魅 法匠 這 堂 課 反 服貿 大 火 風暴 服貿 風暴 風暴 法匠 立法院 之 恥 民主 鬧劇 「 解嚴 後 的 民主 嬰兒 」

Presupposition (PS, 預設立場)：「 占領 國會 」就 能 等同 於「 領導 國家 」嗎 ？「 領導 國家 」 難道 不 需要 經過 公職 選舉 嗎 ？ 」 這 種 說法 不 是 很 有力 嗎 ？ 不 就是 司法 干預 政治 的 危險 嗎 ？ 不 是 嗎 ？ 不 覺得 太過 自大 及 反 民主 嗎 ？ 並 不 令 人 意外 他 不 就 能 把 支持 服貿 的 社會 力量 拉進 來 當成 自己 的 憑藉 嗎 ？ 但 整 個 政府 的 顏面 可以 如此 被 踐踏 嗎 ？你們 心 中 真的 還 有 別人 嗎 ？ 又 豈 對 台灣 有利 ？ 豈 對 台灣 有利 ？台灣 的 政府 體制 可以 伸縮 自如 到 隨便 誰 來 指揮 嗎 ？ 否則 台灣 豈 有可能 訂出 一 紙 「 只 取 不 給 」 的 條約 ？ 國家 能否 由 霸占 國會 者 來 領導 學生 是否 應當 自問 ： 法官 在 擒縱 之間 豈非 更 不符 比例 原則 ？ 竟 法官 難道 是 活 在 「 象牙塔 裡 」 的 人 ？ 絕非 竟 竟 以 可以 救濟 的 「 程序 正義 」 賠上 分秒必爭 且 無可 逆轉 的 「 台灣 全球化 機遇 」 ， 竟 莫非 就是 看穿 了 總統 的 軟弱 可 欺 ？ 被 卡住 不能 動彈 的 議案 豈止 服貿 協議 一案 ？ 豈不 變成 自己 所 鄙薄 的 人 ？ 這 不 是 在 阻斷 台灣 加入 ＴＰＰ 及 ＲＣＥＰ 之 路 嗎 ？ 豈不 這 真的 有 把 服貿 當 回 事 嗎 ？ 這 豈 符合 其 「 捍衛 民主 」 的 口號 ？ 這 豈非 政黨 政治 的 扭曲 ？ 這 難道 不 是 民主 的 失控 ？ 這些 人 有 認真 看待 學生 的 不滿 嗎 ？ 都 不可能 簽名 交出 自己 的 代議權 吧 ？ 難道 應當 鋪紅 地毯 相迎 ？ 顯然

Thought-terminating cliché (TTC, 格言論證)：可想而知 常識 毫無疑問

The Association for Computational Linguistics and Chinese Language Processing

(new members are welcomed)

Aims：

1. To conduct research in computational linguistics.
2. To promote the utilization and development of computational linguistics.
3. To encourage research in and development of the field of Chinese computational linguistics both domestically and internationally.
4. To maintain contact with international groups who have similar goals and to cultivate academic exchange.

Activities：

1. Holding the Republic of China Computational Linguistics Conference (ROCLING) annually.
2. Facilitating and promoting academic research, seminars, training, discussions, comparative evaluations and other activities related to computational linguistics.
3. Collecting information and materials on recent developments in the field of computational linguistics, domestically and internationally.
4. Publishing pertinent journals, proceedings and newsletters.
5. Setting of the Chinese-language technical terminology and symbols related to computational linguistics.
6. Maintaining contact with international computational linguistics academic organizations.
7. Dealing with various other matters related to the development of computational linguistics.

To Register：

Please send application to:

The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

payment： Credit cards(please fill in the order form), cheque, or money orders.

Annual Fees：

regular/overseas member： NT$ 1,000 (US$50.-)
group membership： NT$20,000 (US$1,000.-)
life member：ten times the annual fee for regular/ group/ overseas members

Contact：

Address： The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

Tel.：886-2-2788-3799 ext. 1502 Fax：886-2-2788-1638

E-mail: aclclp@hp.iis.sinica.edu.tw Web Site: http://www.aclclp.org.tw

Please address all correspondence to Miss Qi Huang, or Miss Abby Ho

The Association for Computational Linguistics and Chinese Language Processing

Membership Application Form

Member ID#：________________

Name：__________________________________ Date of Birth：_______________

Country of Residence：___________________ Province/State：___________________

Passport No.：___________________________ Sex: ________________________

Education(highest degree obtained)：____________________________________

Work Experience：____

____

Present Occupation：____

Address：____

____

Email Add：____

Tel. No：____________________________ Fax No：____________________________

Membership Category：☐ Regular Member ☐ Life Member

Date： _____/_____/_____（Y-M-D）

Applicant's Signature：

Remarks： Please indicated clearly in which membership category you wish to register, according to the following scale of annual membership dues：

Regular Member ： US$ 50.- （NT$ 1,000）
Life Member ： US$500.-（NT$10,000）

Please feel free to make copies of this application for others to use.

Committee Assessment：

中華民國計算語言學學會

宗旨：

（一） 從事計算語言學之研究
（二） 推行計算語言學之應用與發展
（三） 促進國內外中文計算語言學之研究與發展
（四） 聯繫國際有關組織並推動學術交流

活動項目：

（一）定期舉辦中華民國計算語言學學術會議（Rocling）
（二）舉行有關計算語言學之學術研究講習、訓練、討論、觀摩等活動項目
（三）收集國內外有關計算語言學知識之圖書及最新發展之資料
（四）發行有關之學術刊物，論文集及通訊
（五）研定有關計算語言學專用名稱術語及符號
（六）與國際計算語言學學術機構聯繫交流
（七）其他有關計算語言發展事項

報名方式：

1. 入會申請書：請至本會網頁下載入會申請表，填妥後郵寄或E-mail至本會
2. 繳交會費：劃撥：帳號：19166251，戶名：中華民國計算語言學學會
   信用卡：請至本會網頁下載信用卡付款單

年費：

終身會員：	10,000.-	（US$ 500.-）
個人會員：	1,000.-	（US$ 50.-）
學生會員：	500.-	（限國內學生）
團體會員：	20,000.-	（US$ 1,000.-）

連絡處：

地址：台北市115南港區研究院路二段128號 中研院資訊所(轉)
電話：(02) 2788-3799 ext.1502 傳真：(02) 2788-1638
E-mail：aclclp@hp.iis.sinica.edu.tw 網址: http://www.aclclp.org.tw
連絡人：黃琪 小姐、何婉如 小姐

中 華 民 國 計 算 語 言 學 學 會
個 人 會 員 入 會 申 請 書

<table>
<tr><td>會員類別</td><td>☐終身 ☐個人 ☐學生</td><td colspan="2">會員編號</td><td colspan="2">（由本會填寫）</td></tr>
<tr><td rowspan="2">姓　　名</td><td rowspan="2"></td><td rowspan="2">性別</td><td rowspan="2"></td><td>出生日期</td><td>年　月　日</td></tr>
<tr><td>身分證號碼</td><td></td></tr>
<tr><td>現　　職</td><td></td><td colspan="2">學　　歷</td><td colspan="2"></td></tr>
<tr><td>通訊地址</td><td colspan="5">□□□</td></tr>
<tr><td>戶籍地址</td><td colspan="5">□□□</td></tr>
<tr><td>電　　話</td><td></td><td colspan="2">E-Mail</td><td colspan="2"></td></tr>
<tr><td colspan="6">申請人:　　　　　　（簽章）

中　華　民　國　　　年　　月　　日</td></tr>
</table>

審查結果:

1. 年費：

   終身會員：　10,000.-
   個人會員：　1,000.-
   學生會員：　500.-（限國內學生）
   團體會員：　20,000.-

2. 連絡處：

   地址：台北市南港區研究院路二段128號 中研院資訊所(轉)
   電話：(02) 2788-3799　ext.1502 傳真：(02) 2788-1638
   E-mail：aclclp@hp.iis.sinica.edu.tw　網址: http://www.aclclp.org.tw
   連絡人：黃琪 小姐、何婉如 小姐

3. 本表可自行影印

The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)
PAYMENT FORM

Name: ___________________________________(Please print)   Date: _______________

Please debit my credit card as follows: US$ ____________

❑ VISA CARD ❑ MASTER CARD ❑ JCB CARD   Issue Bank:__________________

Card No.: ________ -_________-_________ -_________ Exp. Date:____________(M/Y)

3-digit code: _________ (on the back card, inside the signature area, the last three digits)

CARD HOLDER SIGNATURE: ___________________________________

Phone No.: _________________________E-mail: ____________________________________

Address: ____

PAYMENT FOR

US$ _____ ❑ Computational Linguistics & Chinese Languages Processing (IJCLCLP)

Quantity Wanted: ______

US$ _____ ❑ Journal of Information Science and Engineering (JISE)

Quantity Wanted: ______

US$ _____ ❑ Publications:________________________________

US$ _____ ❑ Text Corpora: ______________________________

US$ _____ ❑ Speech Corpora:_____________________________

US$ _____ ❑ Others: ____________________________________

US$ _________ ❑ Membership Fees ❑ Life Membership ❑ New Membership ❑Renew

US$ _________ = Total

Fax 886-2-2788-1638 or Mail this form to:

ACLCLP

% IIS, Academia Sinica

Rm502, No.128, Sec.2, Academia Rd., Nankang, Taipei 115, Taiwan

E-mail: aclclp@hp.iis.sinica.edu.tw

Website: http://www.aclclp.org.tw

中 華 民 國 計 算 語 言 學 學 會
信用卡付款單

姓名: ____________________________(請以正楷書寫)　　日期:：____________________

卡別：❑ VISA CARD　　❑ MASTER CARD ❑ JCB CARD　　發卡銀行：______________

信用卡號：__________-__________-__________-__________　　有效日期：__________(m/y)

卡片後三碼：______________（卡片背面簽名欄上數字後三碼）

持卡人簽名：_______________________________________(簽名方式請與信用卡背面相同)

通訊地址：____

聯絡電話：____________________________E-mail：_________________________________

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT$__________ ❑ 中文計算語言學期刊(IJCLCLP) _______________

NT$__________ ❑ Journal of Information Science and Engineering (JISE)

NT$__________ ❑ 中研院詞庫小組技術報告_____________________

NT$__________ ❑ 文字語料庫 ____________________________________

NT$__________ ❑ 語音資料庫 ____________________________________

NT$__________ ❑ 光華雜誌語料庫1976~2010

NT$__________ ❑ 中文資訊檢索標竿測試集/文件集

NT$__________ ❑ 會員年費：❑續會　　　❑新會員　　　❑終身會員

NT$__________ ❑ 其他: ____________________________________

NT$_______________ = 合計

填妥後請傳真至 02-27881638 或郵寄至:
11529台北市南港區研究院路2段128號中研院資訊所(轉)中華民國計算語言學學會 收
E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

Publications of the Association for Computational Linguistics and Chinese Language Processing

		Surface	AIR (US&EURP)	AIR (ASIA)	VOLUME	AMOUNT
1.	no.92-01, no. 92-04(合訂本) ICG 中的論旨角色與 A Conceptual Structure for Parsing Mandarin -- Its Frame and General Applications--	US$ 9	US$ 19	US$15	_____	_____
2.	no.92-02 V-N 複合名詞討論篇 & 92-03 V-R 複合動詞討論篇	12	21	17	_____	_____
3.	no.93-01 新聞語料庫字頻統計表	8	13	11	_____	_____
4.	no.93-02 新聞語料庫詞頻統計表	18	30	24	_____	_____
5.	no.93-03 新聞常用動詞詞頻與分類	10	15	13	_____	_____
6.	no.93-05 中文詞類分析	10	15	13	_____	_____
7.	no.93-06 現代漢語中的法相詞	5	10	8	_____	_____
8.	no.94-01 中文書面語頻率詞典（新聞語料詞頻統計）	18	30	24	_____	_____
9.	no.94-02 古漢語字頻表	11	16	14	_____	_____
10.	no.95-01 注音檢索現代漢語字頻表	8	13	10	_____	_____
11.	no.95-02/98-04 中央研究院平衡語料庫的內容與說明	3	8	6	_____	_____
12.	no.95-03 訊息爲本的格位語法與其剖析方法	3	8	6	_____	_____
13.	no.96-01 「搜」文解字－中文詞界研究與資訊用分詞標準	8	13	11	_____	_____
14.	no.97-01 古漢語詞頻表（甲）	19	31	25	_____	_____
15.	no.97-02 論語詞頻表	9	14	12	_____	_____
16.	no.98-01 詞頻詞典	18	30	26	_____	_____
17.	no.98-02 Accumulated Word Frequency in CKIP Corpus	15	25	21	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	4	9	7	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	8	13	11	_____	_____
20.	Computational Linguistics & Chinese Languages Processing (One year) (Back issues of *IJCLCLP*: US$ 20 per copy)	---	100	100	_____	_____
21.	Readings in Chinese Language Processing	25	25	21	_____	_____
				TOTAL	_____	_____

10% member discount: ___________Total Due:__________

- **OVERSEAS USE ONLY**
- PAYMENT： ☐ Credit Card (Preferred)
  ☐ Money Order or Check payable to "The Association for Computation Linguistics and Chinese Language Processing " or “中華民國計算語言學學會”
- E-mail：aclclp@hp.iis.sinica.edu.tw

Name (please print): ____________________________ Signature: ____________________________

Fax: ____ E-mail: ______________________________

Address：____

中華民國計算語言學學會

相關出版品價格表及訂購單

編號	書目	會　員	非會員	冊數	金額
1.	no.92-01, no. 92-04 (合訂本) ICG 中的論旨角色 與 A conceptual Structure for Parsing Mandarin--its Frame and General Applications--	NT$ 80	NT$ 100	_____	_____
2.	no.92-02, no. 92-03 (合訂本) V-N 複合名詞討論篇 與V-R 複合動詞討論篇	120	150	_____	_____
3.	no.93-01 新聞語料庫字頻統計表	120	130	_____	_____
4.	no.93-02 新聞語料庫詞頻統計表	360	400	_____	_____
5.	no.93-03 新聞常用動詞詞頻與分類	180	200	_____	_____
6.	no.93-05 中文詞類分析	185	205	_____	_____
7.	no.93-06 現代漢語中的法相詞	40	50	_____	_____
8.	no.94-01 中文書面語頻率詞典（新聞語料詞頻統計）	380	450	_____	_____
9.	no.94-02 古漢語字頻表	180	200	_____	_____
10.	no.95-01 注音檢索現代漢語字頻表	75	85	_____	_____
11.	no.95-02/98-04 中央研究院平衡語料庫的內容與說明	75	85	_____	_____
12.	no.95-03 訊息為本的格位語法與其剖析方法	75	80	_____	_____
13.	no.96-01 「搜」文解字－中文詞界研究與資訊用分詞標準	110	120	_____	_____
14.	no.97-01 古漢語詞頻表 (甲)	400	450	_____	_____
15.	no.97-02 論語詞頻表	90	100	_____	_____
16	no.98-01 詞頻詞典	395	440	_____	_____
17.	no.98-02 Accumulated Word Frequency in CKIP Corpus	340	380	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	90	100	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	75	85	_____	_____
20	論文集 COLING 2002 紙本	100	200	_____	_____
21.	論文集 COLING 2002 光碟片	300	400	_____	_____
22.	論文集 COLING 2002 Workshop 光碟片	300	400	_____	_____
23.	論文集 ISCSLP 2002 光碟片	300	400	_____	_____
24.	交談系統暨語境分析研討會講義（中華民國計算語言學學會1997第四季學術活動）	130	150	_____	_____
25.	中文計算語言學期刊 (一年兩期) 年份：______（過期期刊每本售價500元）	---	2,500	_____	_____
26.	Readings of Chinese Language Processing	675	675	_____	_____
27.	剖析策略與機器翻譯 1990	150	165	_____	_____
			合　計	_____	_____

※ 此價格表僅限國內（台灣地區）使用

劃撥帳戶：中華民國計算語言學學會　　劃撥帳號：19166251

聯絡電話：(02) 2788-3799 轉1502

聯絡人： 黃琪 小姐、何婉如 小姐　　E-mail:aclclp@aclclp.org.tw

訂購者：______________________ 收據抬頭：______________________

地　　址：____

電　　話：______________________ E-mail:______________________

Information for Authors

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) invites submission of original research papers in the area of computational linguistics and speech/text processing of natural language. All papers must be written in English or Chinese. Manuscripts submitted must be previously unpublished and cannot be under consideration elsewhere. Submissions should report significant new research results in computational linguistics, speech and language processing or new system implementation involving significant theoretical and/or technological innovation. The submitted papers are divided into the categories of regular papers, short paper, and survey papers. Regular papers are expected to explore a research topic in full details. Short papers can focus on a smaller research issue. And survey papers should cover emerging research trends and have a tutorial or review nature of sufficiently large interest to the Journal audience. There is no strict length limitation on the regular and survey papers. But it is suggested that the manuscript should not exceed 40 double-spaced A4 pages. In contrast, short papers are restricted to no more than 20 double-spaced A4 pages. All contributions will be anonymously reviewed by at least two reviewers.

Copyright：It is the author's responsibility to obtain written permission from both author and publisher to reproduce material which has appeared in another publication. Copies of this permission must also be enclosed with the manuscript. It is the policy of the CLCLP society to own the copyright to all its publications in order to facilitate the appropriate reuse and sharing of their academic content. A signed copy of the IJCLCLP copyright form, which transfers copyright from the authors (or their employers, if they hold the copyright) to the CLCLP society, will be required before the manuscript can be accepted for publication. The papers published by IJCLCLP will be also accessed online via the IJCLCLP official website and the contracted electronic database services.

Style for Manuscripts: The paper should conform to the following instructions.

1. Typescript: Manuscript should be typed double-spaced on standard A4 (or letter-size) white paper using size of 11 points or larger.

2. Title and Author: The first page of the manuscript should consist of the title, the authors' names and institutional affiliations, the abstract, and the corresponding author's address, telephone and fax numbers, and e-mail address. The title of the paper should use normal capitalization. Capitalize only the first words and such other words as the orthography of the language requires beginning with a capital letter. The author's name should appear below the title.

3. Abstracts and keywords: An informative abstract of not more than 250 words, together with 4 to 6 keywords is required. The abstract should not only indicate the scope of the paper but should also summarize the author's conclusions.

4. Headings: Headings for sections should be numbered in Arabic numerals (i.e. 1.,2....) and start form the left-hand margin. Headings for subsections should also be numbered in Arabic numerals (i.e. 1.1. 1.2...).

5. Footnotes: The footnote reference number should be kept to a minimum and indicated in the text with superscript numbers. Footnotes may appear at the end of manuscript

6. Equations and Mathematical Formulas: All equations and mathematical formulas should be typewritten or written clearly in ink. Equations should be numbered serially on the right-hand side by Arabic numerals in parentheses.

7. References: All the citations and references should follow the APA format. The basic form for a reference looks like

> Authora, A. A., Authorb, B. B., & Authorc, C. C. (Year). Title of article. *Title of Periodical, volume number*(issue number), pages.

Here shows an example.

> Scruton, R. (1996). The eclipse of listening. *The New Criterion, 15*(30), 5-13.

The basic form for a citation looks like (Authora, Authorb, and Authorc, Year). Here shows an example. (Scruton, 1996).

Please visit the following websites for details.

(1) APA Formatting and Style Guide (http://owl.english.purdue.edu/owl/resource/560/01/)

(2) APA Stytle (http://www.apastyle.org/)

No page charges are levied on authors or their institutions.

Final Manuscripts Submission: If a manuscript is accepted for publication, the author will be asked to supply final manuscript in MS Word or PDF files to clp@hp.iis.sinica.edu.tw

Online Submission: https://ijclclp.aclclp.org.tw/servlet/SignInHandler

Please visit the IJCLCLP Web page at http://www.aclclp.org.tw/journal/index.php

International Journal of
Computational Linguistics & Chinese Language Processing
vol.26 no.1 June 2021

Contents

Papers

ISSN: 1027-376X

The Association for Computational Linguistics and Chinese Language Processing