

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.

Special Issue on “Corpus Linguistics and Discourse Annotations”

Guest Editors: Siaw-Fong Chung, Rafal Rzepka and Shih-ping Wang

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
生知天下之至蹟而不
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.27

No.1

June 2022

ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

Advisory Board

Hsin-Hsi Chen
National Taiwan University, Taipei
Sin-Horng Chen
*National Chiao Tung University,
Hsinchu*
Pak-Chung Ching
*The Chinese University of Hong
Kong, Hong Kong*
Chu-Ren Huang
*The Hong Kong Polytechnic
University, Hong Kong*

Chin-Hui Lee
*Georgia Institute of Technology,
U. S. A.*
Lin-Shan Lee
*National Taiwan University,
Taipei*
Haizhou Li
*National University of
Singapore, Singapore*

Richard Sproat
Google, Inc., U. S. A.
Keh-Yih Su
Academia Sinica, Taipei
Chiu-Yu Tseng
Academia Sinica, Taipei

Editors-in-Chief

Berlin Chen
National Taiwan Normal University, Taipei

Hung-Yu Kao
National Cheng Kung University, Tainan

Associate Editors

Chia-Ping Chen
*National Sun Yat-sen University,
Kaoshiung*
Hao-Jan Chen
*National Taiwan Normal University,
Taipei*
Pu-Jen Cheng
National Taiwan University, Taipei
Min-Yuh Day
Tamkang University, Taipei
Lun-Wei Ku
Academia Sinica, Taipei
Shou-De Lin
*National Taiwan University,
Taipei*

Meichun Liu
*City University of Hong Kong,
Hong Kong*
Chao-Lin Liu
*National Chengchi University,
Taipei*
Wen-Hsiang Lu
*National Cheng Kung
University, Tainan*
Richard Tzong-Han Tsai
*National Central University,
Taoyuan*
Yu Tsao
Academia Sinica, Taipei

Shu-Chuan Tseng
Academia Sinica, Taipei
Yih-Ru Wang
*National Chiao Tung
University, Hsinchu*
Jia-Ching Wang
*National Central University,
Taoyuan*
Shih-Hung Wu
*Chaoyang University of
Technology, Taichung*
Liang-Chih Yu
Yuan Ze University, Taoyuan

Executive Editor: Abby Ho

The Association for Computational Linguistics and Chinese Language Processing, Taipei

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

© The Association for Computational Linguistics and Chinese Language Processing

International Journal of Computational Linguistics and Chinese Language Processing is published four issues per volume by the Association for Computational Linguistics and Chinese Language Processing. Responsibility for the contents rests upon the authors and not upon ACLCLP, or its members. Copyright by the Association for Computational Linguistics and Chinese Language Processing. All rights reserved. No part of this journal may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical photocopying, recording or otherwise, without prior permission in writing form from the Editor-in Chief.

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

Contents

Special Issue Articles:

Corpus Linguistics and Discourse Annotations

Preface: Corpus Linguistics and Discourse Annotations..... i
Siaw-Fong Chung, Rafal Rzepka, and Shih-ping Wang,
Guest Editors

Papers

The Uniqueness in Speech: Prosodic Highlights-prompted
Information Content Projection in Continuous Speech..... 1
Helen Kai-yun Chen, and Chiu-yu Tseng

Topic Development and Boundary Cues in Hakka Conversational
Discourse..... 27
Shu-Chuan Tseng, and Hsiao-chien Liu

應用文步分析探究言語行為—以公共政策網路參與平臺提案
文類為例 [A Move Analysis of Communicative Acts in Petition
Text on the Public Policy Participation Network Platform]..... 53
楊惟婷(Wei-Ting Yang), 謝承諭(Chen-Yu Chester Hsieh),
鍾曉芳(Siaw-Fong Chung)

An N-gram Approach to Identifying the Chinese Linguistic
Signals for the Problem-Solution Pattern in Annotated Online
Health News..... 75
Chen-Yu Chester Hsieh, and Yu-Yun Chang

Let Me Finish!—Speech Patterns of Interruptions in Chinese: A
Corpus-based Study on Parliamentary Interpellations on Taiwan... 111
Christian Schmid, and Chia-Rung Lu

以社群媒體語言建構深度學習模型：以「校正回歸」為例
[Constructing a Deep Learning Model Using Language in Social
Media: The Case Study of ‘Retrospective Adjustment’]..... 153
段人鳳(Ren-feng Duann), 邱淑怡(Shu-I Chiu),
劉慧雯(Hui-Wen Liu)

Preface: Corpus Linguistics and Discourse Annotations

Siaw-Fong Chung*, Rafal Rzepka+ and Shih-ping Wang#

Discourse analysis examines patterns of language across texts and considers the relationship between language and the social and cultural contexts in which it is used. Discourse analysis also considers the ways that the use of a language presents different views of the world and different understandings. It examines how the use of language is influenced by relationships between participants as well as the effect the use of language has upon social identities and relations. It also considers how views of the world, and identities, are constructed through the use of discourse (Paltridge, 2012: 2; cf. McCarthy, Matthiessen, & Slade, 2010).

With the advancement of computational linguistic technology, data-driven approaches, deep learning, and large language models now allow the processing of language with limited (or without) human annotated data. The strong connection between “computation” and “linguistics,” originally established for the term “computational linguistics,” has been replaced by machine-learning-based data science to predict linguistic phenomena. Huang and Xue (2019: 492) said the following on humanity, technology, digital language resources, and NLP tools:

The value of language resources [...] lies not in the data itself but in its accessibility and inter-operability, in the quality of annotation and the quality of knowledge discovery tools. In our highly connected future, data is not just king. In fact, data is life. Language resources as the most human-oriented form of data will continue to anchor the link between humanity and technology.

Yet the layers of hidden semantic dimensions beneath human annotation, which are not easy to mirror with automatic classifiers, still play an important role—humans’ judgement of world views, humans’ understanding of conversational flow and purpose, relationships between people, and the “effect the use of language has upon social identities and relations” (cf. Paltridge, 2012: 2)—and have a place in computational linguistics. This view is becoming weaker but still needs to be raised so that human decision-making in the categorization of data can be valued, and re-evaluated, to re-establish cooperation between computation and linguistics.

This idea brought about the birth of this special issue titled “Corpus Linguistics and Discourse Annotations,” which includes six papers from different perspectives on different

* National Chengchi University, Taiwan

E-mail: sfchung@nccu.edu.tw

+ Hokkaido University, Japan

National Taiwan University of Science and Technology, Taiwan

aspects of discourses—from acoustic features, conversational discourse, parliamentary discussions, political discourse, and petition texts to health news. Most importantly, all of the authors of these papers are willing to share part of their annotated data for the purpose of improving research on computational linguistics and to rebuild the strong relationship between linguistics and computational research.

The first paper is on the acoustic correlates of prosodic highlights in continuous speech by Helen Kai-yun Chen and Chiu-yu Tseng. This paper works along the newly found view that perceived prosodic highlights in continuous speech can function alternatively as the projector of key/focal information allocation, in contrast to the long-held claim that key information is predominantly marked by prominence. The study analyzed four diverse Mandarin speech genres (two spontaneous speech and two read-aloud speech samples) in terms of the information-content unit projector, followed by its respective projection.

The second paper is on topic development and boundary cues in Hakka conversational discourse by Shu-Chuan Tseng and Hsiao-chien Liu. The study investigated topic-specific Hakka conversations and suggested a top-down two-level discourse segmentation approach that took into consideration topic maintenance, including topic and subtopic transition boundaries.

The third paper is on move analysis of communicative acts in petition texts on the Public Policy Participation Network Platform by Wei-Ting Yang, Chen-Yu Chester Hsieh, and Siaw-Fong Chung. In this paper, the method of move analysis was applied to the Public Policy Network Participation Platform (Join Platform), which allows citizens to start and support a petition online and voice their opinions regarding public issues.

The fourth paper is on an n-gram approach to identifying the Chinese linguistic signals of the Problem-Solution pattern in annotated online health news by Chen-Yu Chester Hsieh and Yu-Yun Chang. This article reports an exploratory project that combined the annotation of the Problem-Solution textual pattern in online health news and the quantitative and qualitative methods of corpus linguistics to investigate the linguistic features of particular rhetorical moves.

The fifth paper is on speech patterns of interruptions in Chinese, which reports the corpus-based study on parliamentary discussions on Taiwan by Christian Schmidt and Chia-Rung Lu. Verbal interruptions during parliamentary interpellations based on publicly accessible transcriptions provided by the Legislative Yuan in Taiwan was observed, including turn-taking, cues, and speech markers.

The sixth paper is on a case study that constructed a deep learning model using language in social media by Ren-feng Duann, Shu-i Chiu, and Hui-wen Liu. This research used Facebook posts related to the term “retrospective adjustment” in Taiwan as the corpus during the COVID-19 pandemic period. The authors compared manual coding and prediction using a computational model to explain the differences from the perspective of linguistic features.

When “computation” works together with “linguistics,” we get “computational linguistic” studies that are of higher sensitivity to linguistic knowledge.

參考文獻 (References)

- Huang, C-R., & Xue, N. (2019). Digital language resources and NLP tools. In C-R. Huang, J-S. Zhuo, & B. Meisterernst (Eds.), *The Routledge handbook of Chinese applied linguistics*, 483-497. Routledge.
- McCarthy, M., Matthiessen, C., & Slade, D. (2010). Discourse analysis. In N. Schmitt (Ed.), *An introduction to applied linguistics* (2nd ed.), 53-69. Bookpoint Ltd.
- Paltridge, B. (2012). *Discourse analysis*. Bloomsbury.

The Uniqueness in Speech: Prosodic Highlights-prompted Information Content Projection in Continuous Speech

Helen Kai-yun Chen* and Chiu-yu Tseng⁺

Abstract

Recently, it has been identified that perceived prosodic highlights in continuous speech can function alternatively as the projector of key/focal information allocation. This view provides a novel interpretation to the long-held claim that prominence is used predominantly to mark key information and alludes to the significance of information content planning prompted by perceived prominence. Exploring further information content planning and allocation prompted by prosodic highlights, this study focused on the information content planning unit—"projector" (PJR) and its respective "projection" (PJN) (henceforth PJR-PJN units)—across four diverse Mandarin speech genres. Using the corpus linguistic approach and quantitative analyses, the current study conducted acoustic correlates analyses of F0 realization and pause duration, also the calculation of emphasis-attributed weighting scores based on emphasis levels consistently annotated in the speech data. While the main goal of the study was to profile consistent acoustic realizations across the PJR-PJN units, further confirmation of the patterned deployment of information content in continuous speech was verified. Ultimately, the current results foregrounded the underlying mechanism for information prosody and features unique to speech.

Keywords: Continuous Speech and Discourse, Spoken Corpora and Annotations, Information Content Planning and Allocation, Prosodic Highlights-prompted Projection, Emphasis-attributed Weighting Scores, Information-attributed Weighting Scores.

* Language Center, National Central University

E-mail: helenkychen@gmail.com

⁺ Institute of Linguistics, Academia Sinica, Taipei, Taiwan

E-mail: cytling@sinica.edu.tw

1. Introduction

The current study focused on information content planning and allocation, which is initiated by and associated with perceived prosodic highlights in continuous speech. In speech and discourse, one of the keys to communication is in how interlocutors plan “ahead of time” the allocation of focal information in speech production and perception: for speakers, this mostly concerns how they distinctively and effectively allocate key/focal information to facilitate comprehension. On the other hand, listeners are oriented to salient cues in prosodic manifestations, including the ups and downs of the melody, the pace of the speech output, and other perceptually distinctive cues to help pinpoint the most crucial information in the speech flow. It is our belief that to plan and identify information content in the speech context, perceivable prosodic saliency, particularly prosodic highlights, plays a crucial role. For this reason, we chose to concentrate on prominence¹ in speech and how it is incorporated to project information content allocation in this study.

In order to examine perceived prosodic highlights and their roles in information content projection functions within the speech context, we adopted an unconventional approach to discourse prosody. Specifically, instead of incorporating traditional methods to treat prosodic manifestations with certain predefined phonetic or syntactic units and examine only their face value, we took a holistic, top-down approach and paid attention to the role of upper-level discourse associations. To account for prosodic variations in continuous speech, we adopted the recently proposed hierarchical prosodic phrase grouping (HPG) framework as described in Tseng *et al.* (2005) and Tseng (2010). The main justification for resorting to such a framework was that it could better accommodate and account for the unique features in speech data: as a continuous flow of perceivable signals, the composition of discourse prosody can go beyond the mere concatenations of lower-level linguistic units. By incorporating the HPG framework, our goal was to capture features belonging to speech inclusive of prosody for information content planning at the upper level of discourse realization, as opposed to linguistic prosody that is constrained by lower-level units that are grammar based.

This article will report the follow-up acoustic analyses from a recent study that focused on perceived prosodic highlights-projected information content allocation in the speech context (cf. Chen & Tseng, 2021). Through examining diverse speech data that was annotated for the same discourse-level prosody in hierarchical relationships (i.e. using the HPG framework) and tagged for consistently perceived levels of prominence, it was demonstrated that perceived prosodic highlights involve the indexing function for key/focal information allocation, and thus project

¹ In this study we use “prosodic highlights” and “prominence” interchangeably in referring to the same concept of distinctively perceived segments in continuous speech signals. Note that, in this case, prosodic highlights and prominence used here are not the same as word-level stress and prominence.

information content planning at higher discourse-level prosody (Chen & Tseng, 2021). Based on the same set of speech data and annotation systems, this extended study will report further results of the acoustic analyses of prominence-prompted information content projection units (cf. Chen & Tseng, 2021). Ultimately, the goal was to foreground prominence-correlated information content planning through discourse-level prosody realizations and demonstrate how patterns and features were eventually be derived from “speaking” (i.e., the “parole” in de Saussure, 1966). In the end, we were able to derive prosody specifically for information content planning from speech and discourse, which went beyond seemingly random linguistic prosody in its surface values and realizations.

1.1 Discourse Prosody and Information Content Planning

Prosody, a unique feature of speech, has posed a major challenge in relevant studies concentrating on discourse perception and production. Given that discourse production can go beyond more than just a sequence of sentences (cf. Swerts & Geluykens, 1994) and that continuous speech happens in a highly spontaneous context and is unplanned, how to capture speech prosody in its highly variant realizations is crucial. In most cases, the prosodic realizations of speech are considered and processed by sound units that are segmented by their meta forms or, at most, from units that are syntactically predefined. This follows earlier studies and the long tradition of examining continuous speech signals through syntactic prosody (cf. Lehiste, 1970; see also Cutler *et al.*, 1997 for a review on prosody of spoken languages).

With regard to the allocation of (focal) information by the prosodic realizations in discourse, oftentimes the discussions in relevant literature have focused on focal/new information is directly marked by prominence, for example, pitch accent (such as that conventionally annotated as H*; Silverman *et al.*, 1992; see also Halliday, 1967; Pierrehumbert & Hirschberg, 1990; Watson *et al.*, 2008). However, to associate a high pitch accent with focal information, the pitch accents are aligned with word-level stresses in most cases. In other words, the corresponding unit for prominence realizations are held at the lexical level. On the other hand, Swerts and Geluykens (1994) examined the role of prosody in the structuring of information (i.e., the topic flow and topic changes) and investigated prosodic variables including intonation and pause. In their experiment, the relative pitch height and pause length were associated with information flow markers (Swerts & Geluykens, 1994). Although the speech used in their study was spontaneous, the elicitation of the speech data was controlled for the design and purpose of the experiment.

The recently proposed HPG framework for discourse prosody in continuous speech by Tseng *et al.* (2005) was suggested as an alternative approach (see the additional explanations in Tseng & Su, 2008; Tseng, 2010, 2013). The main strength of the HPG framework is that it is not text-bounded, nor is the relationship between discourse-prosodic units (DPUs)

predetermined grammatically. Instead, the target is how continuous speech signals can be processed from a global viewpoint. According to Tseng (2013), the merit of adopting the HPG framework is to purposely distance it from the possible connotations associated with lower levels of linguistic information, while foregrounding the contribution to higher discourse-level prosody, which also includes discourse-paragraph associations and information content planning. The HPG framework has been adopted in several recent studies focusing on the prosodic features of higher discourse levels in various continuous speech genres (cf. Tseng & Su, 2012, Chen *et al.*, 2016).

1.2 Prosodic Highlight Prompted Information Content Projection

In an exploration of perceived prosodic highlights as an index of information content planning and projection, Chen *et al.* (2016) and Chen and Tseng (2021) reported the analyses of perceived prominence that was consistently annotated across continuous speech. Based on the data from four diverse speech genres that were preprocessed and annotated using HPG, the studies established two information-content indices prompted by prosodic highlights. It was demonstrated that far more tokens of prosodic words with perceivable prominence tags were incorporated into speech to forecast, ahead of time, speech planning and to “project” the allocation of focal information. For instance, in the following examples, the emphasis marked *zuizao de yipian* ‘the earliest entry of’ projects the following noun phrase (NP) *wenzhang* ‘the article’ in (1), whereas in (2), the perceived prominence-indexed *tixing nin* ‘to remind you’ projects the following clause, which contains key information such as *ziwaixian* ‘UV rate’ and *guoliangji* ‘extreme level’²:

(1)

L: 那也是/最早的一篇/文章.

Na yeshi /zuizao de yipian/ wenzhang.

that also.COP earliest DE a.CL article

‘That is also the earliest entry of the article.’ (Chen & Tseng, 2021: 197)

² The concept of prosodic highlights-prompted information projection is shown in the following example from Goodwin (1996: 372), in that the enhanced intonation from a specific part of the discourse is interpreted as projecting focal information:

(i) Nancy: Jeff made en asparagus pie

It was s : so: goo:d.

Tasha: I love it.

According to Goodwin (1996), the prominently pronounced adverb “so” can be interpreted as a projector of the next bit of interaction, as it serves as a kind of prompt for the following adjective “good.” In this example the adjective “good” serves as the main predicate, providing focal and possibly new information.

Prosodic Highlights-prompted Information Content Projection in Continuous Speech

(2)

S: 特別/提醒您/目前白天紫外線都是過量級.

tebie	/tixing nin/	muqian	/baitian/	ziwaixian
especially	remind 2SG	at.the.moment	day.time	UV rate
doushi	/guoliangji/. (WB)			
all.COP	extreme level			

‘Please be reminded especially that at the moment the UV rate during the daytime has reached the extreme level.’

As suggested in Chen and Tseng (2021), in addition to directly marking new/focal information, it has been found that prosodic highlights in continuous speech can be incorporated to index “specific parts of discourse” (e.g., Falk, 2014:8), and thus function to orient listeners’ attention to focal information allocation in speech production. The advantage of incorporating such information projection prompted by prominence, according to Chen and Tseng (2021), is to help eliminate potential prediction errors in speech perception (i.e., Clark, 2013; Auer, 2015; Dille, 2016) and hence facilitate successful communication.

With the assumption that the allocation of prosodic highlights directly reflects the deployment of information content in speech, Chen and Tseng (2021) conducted relevant analyses concentrating on the information content planning unit—“projector” (PJR) and its respective “projection” (PJN) (henceforth **PJR-PJN** units)—prompted by perceived prominence. It was demonstrated that while planning for prosodic highlights-prompted projection, speakers in general were oriented toward a “heavy-to-light” information-attributed weighting scores distributed across the PJR-PJN units (Chen & Tseng, 2021). The results showed that the prosodic highlights-correlated information content planning was realized in a fixed pattern. The main contribution of the study was clarifying that prosody-attributed information content planning in continuous speech takes place at a specific discourse-prosodic level based on the HPG framework (Chen & Tseng, 2021).

1.3 The Current Study: A Preview

The current study was a sequel to the findings on prosodic highlights-initiated PJR-PJN units reported in Chen and Tseng (2021). Following from the assumption of the direct correlation between perceived prosodic highlight distribution and information content allocation, this article will report further analyses based on the PJR-PJN units consistently annotated in data from continuous speech. With the same set of speech data annotated by the corresponding discourse-prosodic levels based on the HPG framework, and according to the same perceived prominence-level annotations, this extended study extracted acoustic measurements from the

PJR-PJN units. In addition, the results provided further validation of the calculation of information-attributed weighting scores across the PJR-PJN unit (cf. Chen & Tseng 2021). Based on the results, the findings suggested that there was an extended substantiation of prosody-attributed information content planning, especially at the higher level of DPUs in continuous speech.

In the present analyses, we chose to concentrate particularly on acoustic correlates across the PJR-PJN units, including a) F0 realization³ and b) pause duration, among the possible acoustic correlates.⁴ As for the validation of the emphasis-attributed weighting scores distribution across the PJR-PJN units, further statistical analyses were carried out. The main difference between Chen and Tseng's (2021) previous study and the results reported in this paper is mainly in that we included the PJR-PJN units in the projection trajectories of various sizes. Although the sizes of information content planning and projection differed from case to case, our analyses still demonstrated identifiable patterns of acoustic realizations and distributions of information-attributed weighting score. As will be shown, the results pinpointed information content planning in correlation with advance prosody prompting, not only in a patterned F0 contour but also in longer pause and heavier information loading that were required at the initiation of the PJR-PJN units. We believe that the results are significant in demonstrating the role of advance prosodic prompting in information projection. The current results will shed light on information content planning in online speech production, and the establishment of information prosody that is unique in speech.

2. Speech Data and Annotations

2.1 Speech Data

Continuous speech data in Taiwanese Mandarin from four diverse genres were incorporated for the purpose of the present analyses. Of the four speech genres, two were spontaneous speech and the other two were read speech. One of the two spontaneous speech genre was a university classroom lecture (henceforth SpnL), taught and delivered by a male professor (i.e., Tseng *et*

³ Some of the preliminary observations regarding F0 realization throughout the PJR-PJN units have been reported earlier in Chen *et al.* (2016).

⁴ Although we chose to focus on the acoustic cues of intonation (F0) and pause, this does not mean that other prosodic cues are irrelevant. The justification for concentrating on only these two acoustic correlates was mainly that each PJR-PJN unit identified was an independent case and had different length (ranging from one prosodic phrase to three prosodic phrases; see the results in Section 4.1.). Since each PJR-PJN was an independent unit, other cues (such as final lengthening at the end of PJR-PJN units) were not the focus of the current analysis. Moreover, with regard to amplitude, given that the sizes of the PJR-PJN units differed case by case, we assumed that it would be difficult to generate consistent amplitude results from the tokens identified.

Prosodic Highlights-prompted Information Content Projection in Continuous Speech Speech

al., 2008), whereas the other speech genre was a spontaneous informal interaction (SpnC) taken from a corpus of face-to-face interaction in Taiwanese Mandarin (Chen *et al.*, 2012). The read speech, on the other hand, included data from the tasks of prose reading (CNA) and weather broadcast simulations (WB), both of which were culled from the Sinica COSPRO corpus (Tseng *et al.*, 2003; Tseng *et al.*, 2005). Note that we incorporated speech data from different genres for the purpose of comparing features that belong to read speech and spontaneous speech/discourse. Table 1 summarizes the total duration of the data from each speech genre, with additional information on the equivalent number of syllables:

Table 1. Summary of total time and number of syllables in the data from four speech genres

Corpora/ Genres	Total Time (min.)	Total Number of Syllables
SpnL	145	33,306
SpnC	54	10,756
CNA	50	22,988
WB	28	14,083

Although the total duration of each genre differed and was not balanced across speech genres, we ensured that there were ample acoustic features present in the target annotated tokens, especially for the purpose of the current acoustic analyses.⁵

2.2 Data Preprocessing and Annotations

First, the selected speech data underwent automatic preprocessing of force alignments using the HTK Toolkit. The output was followed-up by manual spot-checking and then adjusted by the trained transcribers. The next step of data preprocessing involved the annotations of prosody-related information in independent layers. These tasks were carried out by experienced annotators⁶ who tagged the data for the following information: (i) level of DPUs; (ii) level of perception-based prosodic highlights; and (iii) information content planning PJR-PJN units (cf. Chen & Tseng, 2021).

⁵ As the current speech data were taken from six different speakers (three male and three female speakers) in total, in the following acoustic analyses we normalized the measurements in order to avoid the problem of speaker idiosyncrasy.

⁶ The “experienced taggers” (and trained transcribers) in this study were annotators who had undergone preliminary training for at least three to six months. After the training, these annotators continued working with the same data for at least one year. When working on each annotation task, they had to reach a minimum level of consistency rate from the initial training of a certain task before continuing on (see also the sections on annotations to follow).

2.2.1 Annotation Scheme for Discourse-Prosodic Units (DPU)

We first annotated all the speech data for prosody-based breaks and boundaries following the framework of the HPG framework, according to which, five DPU levels with hierarchical relationship were distinguished, and these were marked B1 through B5, corresponding respectively to syllable (SYL), prosodic word (PW), prosodic phrase (PPh), breath group (BG) and multiple phrase speech paragraph (PG). Beyond the lexicon-based and grammar-correlated PW and PPh levels in the HPG framework, there were two more higher-level units and one was at the BG level, which was defined as a physio-linguistic unit constrained by a change of breath while speaking continuously (cf. Lieberman, 1967; Tseng, 2010). As for the highest-level PG, it was mostly discourse based and was predominantly defined by major topic changes. By default, the boundary breaks, prosodic units, and their relationships within the hierarchy were accounted for as follows: SYL/B1 < PW/B2 < PPh/B3 < BG/B4 < PG/B5 (cf. Tseng, 2010).

In the current study, the annotation of the DPUs was carried out by marking boundary breaks in hierarchical relationships, instead of predetermined by any type of lexical or syntactic relationship. To ensure that the annotations reached a certain level of consistency, the participating annotators⁷ had to reach at least an 80% consistency rate during the initial training to continue the task. During and after the annotation process, both intra- and inter-annotator consistency were constantly checked to ensure the agreement was reached and accuracy maintained at a level of least 95% agreement among the annotators.

2.2.2 Annotation Scheme for Perceived Prosodic Highlights

In a separate layer, all the speech data were additionally annotated for perception-based emphasis and non-emphasis tokens (ETs/non-ETs). Following the definition described in Tseng *et al.* (2011) and Tseng (2013), this annotation scheme for perceived prominence was marked by strength levels, from reduction to the most emphasized, and divided into four relative degrees, defined respectively as follows:

- E0 -- reduced pitch, lower volume, and/or contracted segments
- E1 -- normal pitch, normal volume and clearly produced segments
- E2 -- raised pitch, louder volume and irrespective of the speaker's tone of voice
- E3 -- higher raised pitch, louder volume, and with a change in the speaker's tone of voice

The rationale behind adopting this scheme for annotating prominences was based on the belief that only a limited number of contrastive degrees can be consistently perceived while processing

⁷ At least ten annotators participated in the task of annotating the DPUs in the current speech data.

Prosodic Highlights-prompted Information Content Projection in Continuous Speech

continuous speech.⁸ In the annotation of perception-based prominence, the trained annotators simply tagged the speech data in a string that consisted of ETs (i.e., E2 and E3) and non-ETs (i.e., E0 and E1).⁹ Among the four speech genres, only spontaneous speech (i.e., SpnL and SpnC) was tagged for the additional level of reduction (E0), as we assumed that speakers rarely carried out reduction in reading tasks.

For the annotations of perceived prominence, at least eight annotators¹⁰ were involved in the task. In order to carry out the reliability check, we first assigned one to two “reliable” annotators who were more sensitive to prominence-level distinctions. Their tagging results were considered the “gold standard.” As for the rest of the annotators, they had to reach at least an 80% agreement level compared with the reliable annotators’ tagging results, to continue with the task. For the final annotation, the accuracy level had to reach at least 95% of agreement among the annotators.

2.2.3 Identification of Information Content Planning Units

The information content planning PJR-PJN units were annotated via a separate task in yet another independent layer. First, we started with the identification of the prosodic highlights-prompted PJR. The identification of the PJR index was based on the ETs (i.e., E2 and E3) that had already been annotated in the current data. Each E2 and E3 were broken up by a PW unit. Following the principles of categorizing prominence-prompted information content planning proposed by Chen and Tseng (2021), the PJR units were instances in which the speakers incorporated emphasis in a particular PW unit to head-up the deployment of key information in speech planning. In the following examples, the speech strings in between the slashes are the PW units with an E2 prominence level tagged under the current annotation scheme. In (3), the PW unit *bingbu zhidao* ‘not (really) know’ is categorized as a PJR unit. Moreover, in (4), which

⁸ Since the annotation of prominence levels was mainly perception based, the annotators were not given specific instructions to correspond a prominence level to any absolute acoustic value (i.e., they were never given the instruction that an E2 tag would equal a fixed range of F0 measurements in number). We wanted the annotation of prominence to closely and faithfully reflect the perception of the speech signals. Moreover, given that the level of contrast degree was limited, in general the annotators working with this annotation scheme did not have much difficulty in deciding, for example, a two-way distinction between E1 and E2.

⁹ Since in Mandarin the language does not actually carry pitch accent at the word level, our annotation scheme was distinguished from the model of prosody-related prominence proposed by Kohler (1997) and the framework discussed in Baumann *et al.* (2016), in that the current tagging scheme for prominence level was not lexically based nor syntactically predefined.

¹⁰ Some of the annotators who worked on the DPU annotation also worked on the task for prominence-level annotation. However, those annotators did not work on the two tasks simultaneously. In other words, they trained for the two tasks and worked on each separately.

is repeated from (2), the PW unit *tixing nin* ‘to remind you’ is identified as a PJR unit.

(3)

- L: 中文是, 中文的文字是一堆字. 那麼你/並不知道/哪裏是一個詞.
 Zhongwen shi zhongwen de wenzi shi yidui zi.
 Chinese COP Chinese DE text COP a.CL character
 Name ni /bingbu zhidao/ nali shi yige ci. (SpnL)
 then 2SG not know where COP a.CL lexical.word
 ‘(As for) Chinese, the texts in Chinese are presented as a bunch of characters. Thus, you don’t really know which part equals a word.’

(4)

- S: 特別/提醒您/目前白天紫外線都是過量級.
 tebie /tixing nin/ muqian /baitian/ ziwaixian
 especially remind 2SG at.the.moment day.time UV rate
 doushi /guoliangji/. (WB)
 all.COP extreme level
 ‘Please be reminded especially that at the moment the UV rate during the daytime has reached the extreme level.’

Following the identification of PJR units, we turned to the delineation of the respective PJN units, which were identified as anticipated syntactic/semantic/prosodic completion whose trajectory covered at least a piece of focal information (cf. Chen & Tseng, 2021). As suggested by the discussion of prosodic-highlights prompted projection in Chen and Tseng (2021), the projection trajectory of each PJR unit was realized by a different size, from the local to the global. The current study adopted the similar term “projector-projection” (PJR-PJN) coined in Chen and Tseng (2021: 197) to refer to the prosodic highlights-indexed PJR unit, which was followed immediately by the respective PJN unit. Two additional examples below illustrate the PJR-PJN units by the proposed definition:

(5)

- L: 那也是/最早的一篇/文章.
 Na yeshi /zuizao de yipian/ wenzhang. (SpnL)
 that also.COP earliest DE a.CL article
 ‘That is also the earliest entry of the article.’ (Chen & Tseng, 2021: 197)

Prosodic Highlights-prompted Information Content Projection in Continuous Speech

(6)

L: /為什麼直-/直接比對/字也有/困難?因為我們的/詞的/結構是非常 flexible 的。
 /Weisheme zhi-/ zhijie bidui /zi ye you/ kunnan?
 why di- direct match word also have difficulty
 Yinwei women de /ci de/ jieyou shi
 because 1PL DE lexical word DE structure COP
 feichang flexible de. (SpnL)
 quite flexible DE
 ‘Why is there difficulty in matching words directly? (It is) because the composition of the word structure is quite flexible.’ (Chen & Tseng, 2021: 197)

In (5), which is repeated from (1), the prosodic highlights-prompted PW unit *zuizaode yipian* ‘the earliest entry’ as a PJR unit has a respective projection trajectory ending with the NP *wenzhang* ‘article’, as explained earlier. Turning to (6), the prosodic highlights-indexed PW unit *weishenme* ‘why’ is also categorized also as a PJR unit, and the prosodic highlights-prompted PJR unit entails a projection, with its trajectory extending to the end of the following clause which is initiated by the connective *yinwei* ‘because’. According to the definition by Chen and Tseng (2021), the PJN unit’s trajectory in (6) covers at least one piece of focal information (including examples such as *zi* ‘character’ and *ci* ‘lexicon’ and the foreign word ‘flexible’). Hence both (5) and (6) demonstrate that the PJR-PJN units are of different sizes, from the immediate local projection (as shown in [5]) to the more global one (as shown in [6]). Figure 1 presents an illustration of the annotation for (6) taken from Chen and Tseng (2021), inclusive of the DPU levels and prosodic highlight annotations:

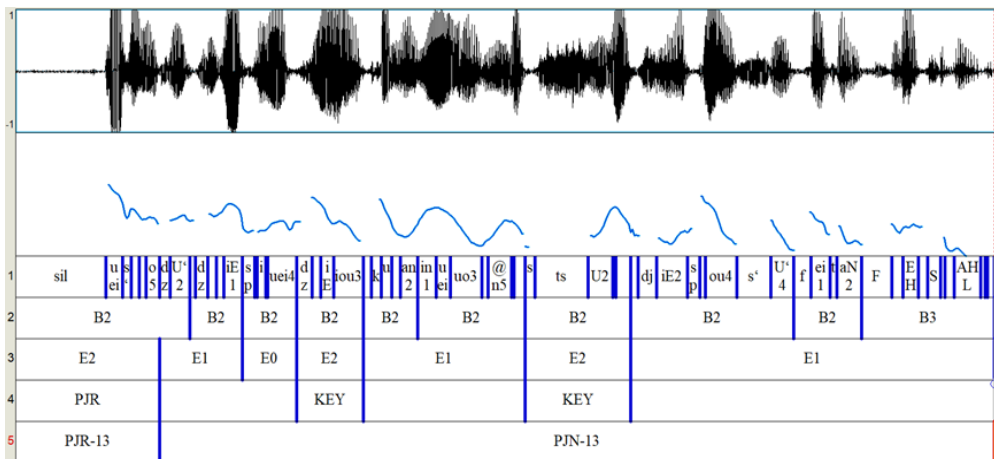


Figure 1. Illustration of the annotation schemes for the DPU levels (in the second layer beneath the spectrogram), prominence levels (in the third layer beneath the spectrogram), and PJR-PJN units using PRAAT (Boersma & Weenink, 2015)

Finally, in terms of annotation consistency rate checking, the identification of prosodic highlights-prompted PJR units was carried out by at least five annotators.¹¹ For the categorization of the PJR tags, the results had to reach 80% agreement among the annotators, and then the PJN trajectory of each PJR instance was demarcated. The annotators checked and discussed each case separately until a final consensus on the trajectory range of each PJN unit was reached.

3. Methodology

3.1 Acoustic Features Extraction

The methodology incorporated in the current analyses involved mainly the extraction of acoustic features, including F0 and pause duration, among other acoustic correlates. First, F0 values (in semitone) across the PJR-PJN units were automatically extracted using the software program PRAAT (© Boersma & Weenink, 2015). In order to facilitate further comparison and eliminate factors from speakers' discrepancy and idiosyncrasy, all the extracted F0 values were subjected to Z-score normalization. Then the next step was to calculate the average F0 values derived from the sampling points, including (i) the PJR at the initiation of the prosodic highlights-prompted projection; (ii) the ending PW at the completion of the PJN; and (iii) the PW units at the pre-/post-PPh boundaries, depending on the trajectory size of the projection (by PPh unit). Figure 2 illustrates the sampling points of a PJR-PJN unit:



Figure 2. Illustration of F0 sampling points of a PJR-PJN unit with a projection of three-PPh units

After deriving the average F0 values, we further attempted the removal of the intonation effect from the higher-level DPUs. This was carried out by remodeling the F0 slope based on PPh units, via turning the value of the F0 slope into 0.

For pause duration, we extracted the duration of silent pauses (in millisecond) located in the following positions: (i) the initiation of the PJR, which was defined as from the off-set of

¹¹ For this annotation task, the annotators included the first author of the paper. As for the other annotators, they had also worked on the DPU and prominence-level annotation tasks. Hence all annotators were quite familiar with the annotation scheme.

Prosodic Highlights-prompted Information Content Projection in Continuous Speech Speech

the PW unit immediately preceding the PJR to the onset of the PJR; and (ii) the initiation of the PJN, which was defined as from the off-set of the corresponding PJR to the onset of the PJN. After the pause durations were derived, we further obtained the mean values of the pause durations in both positions.

3.2 Emphasis-attributed Weighting Scores Calculation

To calculate the emphasis-attributed weighting scores, we followed a similar rationale for modeling prominence-correlated distribution of information-attributed weighting scores proposed by Tseng (2010) and Chen and Tseng (2021) and assumed that there was a direct association between the levels of perceived emphasis annotations and information-attributed weighting scores. Adhering to this assumption, the weighting scores were arbitrarily assigned by using the following formula:

$$(7) \quad \text{Score}(t_n) = \begin{cases} 0, & \text{if label} = E0 \\ 0, & \text{if label} = E1 \\ 1, & \text{if label} = E2 \\ 2, & \text{if label} = E3 \end{cases}$$

In the formula above, the t represents each ET annotated across the current speech data. One additional note is that, as explained in Section 2.2.2, in annotating the perceived prominence degrees of the current spontaneous speech data, the SpnL and SpnC were both tagged with one extra level of reduction (E0). In order to calculate the information-attributed weighting scores on the basis of the same set of prominence levels, initially we merged the E0 tags with the E1 tags in the SpnL and SpnC and assigned a score of 0 to both.¹²

After the scoring assignment, we calculated the average information-attributed weighting scores across the PJR-PJN units by PW units and averaged the weighting scores derived from each PW within the PJR-PJN units, which ranged from one to three PPh units according to the projection trajectory size. Finally, we conducted correlation analysis to examine the relationship between the average weighting scores and the PJR-PJN units with different trajectory sizes.

¹² Initially we merged E0 and E1 tags for the calculation of the weighting scores purely for the purpose of comparing the current read speech and spontaneous speech data with the same set of prominence levels. We also attempted the further calculation of contrast degree by acoustic cues (including F0, duration and intensity) between all the E0 and E1 tags from the SpnL and SpnC. It was found that all the acoustic features were significantly distinctive in the SpnC data, while for SpnL only the duration feature was distinguished. Hence in the analysis reported later in the paper, we further manipulated the E0 tags from spontaneous speech by assigning a score of -1 to all the reduction tags (see section 4.4.2).

4. Acoustic Profiles and Emphasis-attributed Weighting Scores of the PJR-PJN Units

This section will present the analyses of the acoustic realizations and the results of emphasis-attributed weighting scores derived from the information content planning PJR-PJN units. For the acoustic profiles, we focused on the realization of intonation contours throughout the PJR-PJN units in F0 and pause duration in correlation with the initiation of the PJR and PJN. In addition, we examined the correlation between the emphasis-attributed weighting scores and the projection trajectory size of the PJR-PJN units, which shed light on the overall distribution and planning of information content that was prompted by the prosodic highlights.

4.1 Calculation of PJR-PJN Units by PPh Units

Before the analyses, we took an initial step to examine the general distribution of the PJR-PJN unit across the speech data from the four different genres. As suggested previously, the trajectory size of the projection varied for each PJR-PJN unit (Chen & Tseng, 2021). It was thus essential to first identify the projection range distribution of all the PJR-PJN units. As shown in the results from Chen *et al.* (2016) and Chen and Tseng (2021), it was found that over 90% of the PJR-PJN units were accounted for by up to three PPh units. With the identification of a PPh unit as the basic planning DPU for the PJR-PJN units, we further calculated the total number of PJR-PJN units by PPh units compared with the total number of PPh units across the four speech genres. The results shown in Table 2 provide a further illustration of the proportion of PJR-PJN and PPh units in each speech genre:

Table 2. Summary of the total number of syllables in the PJR-PJN units by PPh units and the total number of PPh units in the four speech genres

Corpora/ Genre	Total Number of Syllables in the PJR-PJN Units by PPh Units	Total Number of PPh Units
SpnL	1,257 (28%)	4,535
SpnC	347 (25%)	1,372
CNA	821(48%)	1,702
WB	324 (38%)	861

Hence in the following analyses, we adopted the PPh unit as the base planning unit to estimate the acoustic correlates and weighting score calculation of the PJR-PJN units. To extend further the findings from Chen and Tseng (2021), we included the PJR-PJN units with projection trajectories ranging from one to three PPh units in the current data.

4.2 Acoustic Correlate: F0 Realizations (with and without Intonation Effect)

Following the methodology described in Section 3.1, we calculated the mean F0 values by the PW units at each sampling points, including the initial and final PW of the PJR-PJN units, as well as the PW units by the PPh boundaries in each PJR-PJN unit, and the results are summarized in Figure 3. On the other hand, Figure 4 presents the results of the F0 measurements at the same sampling points after removing the intonation effect from the higher-level DPUs. Note that both figures present the results according to the trajectory size of the PJR-PJN units, from one up to three PPh units.

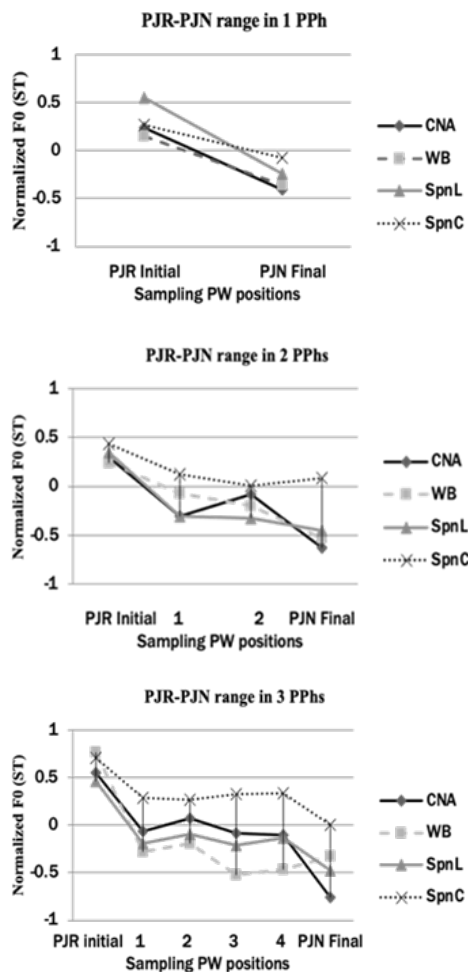


Figure 3. F0 of each PJR-PJN unit (sampling points by position: 1= PW prior to first PPh boundary; 2 = PW after first PPh boundary; 3 = PW prior to second PPh boundary; 4 = PW after second PPh boundary)

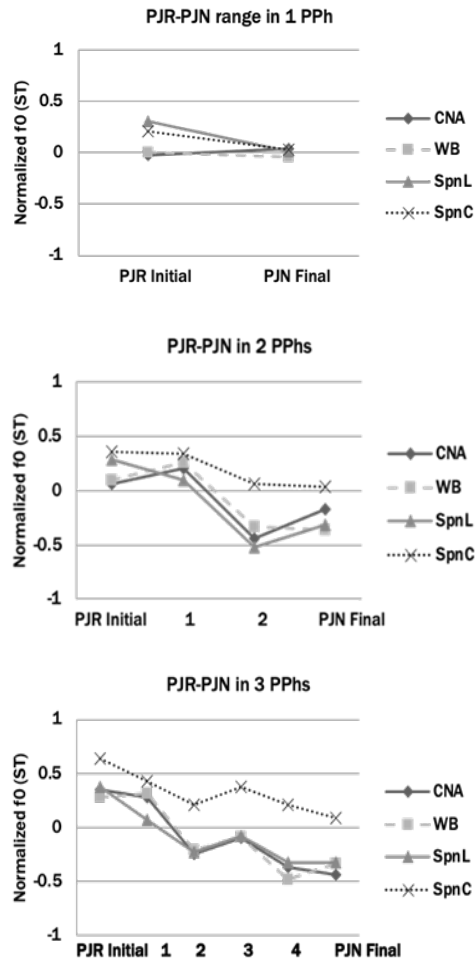


Figure 4. *F0 of each PJR-PJN unit without intonation effect (sampling points by position: 1= PW prior to first PPh boundary; 2 = PW after first PPh boundary; 3 = PW prior to second PPh boundary; 4 = PW after second PPh boundary)*

4.2.1 Results

First, a general tendency of a “high-to-low” pitch contour was observed across the PJR-PJN units as shown in the three panels of Figure 3. This falling contour was noticeable, regardless of the projection size. Although there were occasional exceptions when a slight rising contour was observed in the PJR-PJN units, (i.e., in the CNA data), when an information content unit extended to two PPh units, the rising contour never reached a point higher than the F0 derived from the initial point of the corresponding PJR unit. Slight final-rising contours were also

Prosodic Highlights-prompted Information Content Projection in Continuous Speech

observed in the SpnC data with two PPh units and in the WB data when the unit expanded to three PPh units. However, the final rising contours in both cases never reached a point higher than the F0 values extracted from the corresponding PJR initiation point. Above all, we found that the F0 values derived from the beginning of PJR units and the ending of PJN units were distinguished, regardless of the trajectory sizes. Further statistical tests indicated that significant differences were present ($h=1, p<0.05$ across all three panels in Figure 3) and thus substantiated the observation of the general falling intonation contour across the PJR-PJN units.

To further validate the falling contours observed, we attempted the removal of the intonation effect from the higher-level DPUs. As presented in Figure 4, after removing the intonation effect, the falling pitch contour was still sustained. Even though there were also slight rising contours both within the projection trajectories and at the end of the projection trajectories in some of the data, the rising contours did not reach a point higher than the F0 values in the corresponding initial PJR units. The only noticeable exception was in the read speech genres (i.e., WB and CAN), in which the PJR-PJN units equaled one PPh unit. *T*-test results also confirmed that the F0 values of the beginning of the PJR units and the ending of the PJN units were distinguished, (all $h=1, p\leq 0.05$), except for instances in which the projection trajectory was local and within one PPh unit in the read speech genres.

4.2.2 Discussion

The results above demonstrated that, when planning for prosodic highlights-prompted PJR-PJN units as the information content planning units, in general the speakers initiated the intended information content planning units from a higher F0 and continue with a gradual falling contour across the projection trajectories. Although there were cases in which slight rising contours were observed, the rising pitch never reached a point higher than the F0 values derived from the beginning of the PJR units. Furthermore, a general tendency was observed in that, the larger the projection was (i.e., when the trajectory expanded over two PPh units), the greater the difference between the mean F0 values from the beginning of the PJR units and the ending of the PJN units was. This in turn reflected that in the fore-planning of larger information projections, the speakers had to prepare to start the PJR unit at a higher F0 to allow for the further manipulation and allocation of the prosodic variations within the planned projection trajectory.

After removing the intonation effect from the higher-level DPUs, the falling pitch contour across the PJR-PJN units was still maintained. Interestingly, when the projection size was only within one PPh unit and of local planning, the falling contour was not as obvious: the F0 values of the projection trajectories of the initiations and endings were barely distinguishable. In the end, it was only when we considered the global projection of information content that the falling contour was of distinctive significance. Chen *et al.* (2016) reported their results from further calculations of the down-stepping degrees across the PJR-PJN units of different trajectory sizes,

and a positive correlation between the down-stepping degrees and the projection trajectory sizes was identified. As shown in Table 3 repeated from Chen *et al.* (2016), the longer the projection trajectory size was, the larger the degree of differences derived from the beginning of the PJR units and the end of the PJN units was:

Table 3. Down-stepping degree across the PJR-PJN units, calculated by PPh units (Chen *et al.*, 2016)

Down-stepping Degrees across PJR-PJN Units			
Genre	Within a PPh	Across 1 PPh	Across 2 PPhs
CNA	0.067	0.234	0.789
WB	0.049	0.452	0.614
SpnL	0.294	0.600	0.700
SpnC	0.173	0.316	0.553

According to Chen *et al.* (2016), the result from the down-stepping degree calculations further reinforced that the overall intonation planning across the PJR-PJN units was not due to the influence of the higher-level intonation effect. In other words, in the actual planning of the information content within a larger projection trajectory that was prominence-prompted, the speakers resorted to a noticeable falling contour and a larger down-stepping degree. This was for the purpose of accommodating more variations in the prosodic highlight allocations within the projection trajectories to reflect focal information allocations.

4.3 Acoustic Correlate: Pause Duration

The second acoustic feature we turned to was the pause duration. In particular, we focused on the duration of silent pauses located prior to the initiation of a PJR unit and in between PJR and PJN units for planning the projection trajectories. Similar to the findings on pause durations in topic flow in spoken discourse by Swerts and Geluykens (1994), it was hypothesized that the longer the PJR-PJN unit was, the more time required for the speaker to initiate the prosodic-prompted PJR unit and plan the projection trajectory; hence, the longer the silent pause duration prior to the initiation of both the PJR and PJN units. Here we focused on the estimation of the correlation between the average pause duration and size of the PJR-PJN units (by PPh unit).

4.3.1 Results

As demonstrated in Figure 5, a general tendency was observed in that, when the projection trajectory size increased, the pre-PJR pause duration was also longer. This was most obvious when comparing PJR-PJN units in the one to two PPh units range. Turning to the PJR-PJN units

Prosodic Highlights-prompted Information Content Projection in Continuous Speech

in the three PPh units range, there were exceptions from the speech genres of WB and SpnL, when the pre-PJR pause was slightly shorter than the average pause duration preceding the PJR-PJN units in the two PPh units range. To verify, we further performed *t*-tests between the average pause duration derived from the PJR-PJN units in the one to three PPhs range, and the results indicated that significant differences were found in the data from both read speech genres (i.e., CNA and WB, both $h=1, p<0.05$). As for the pre-PJN pause duration, the results shown in Figure 6 revealed a similar tendency in that the larger the projection size (i.e., up to three PPh units), the longer the pre-PJN pause was. Further statistical results also showed significant differences in the average pre-PJN pause duration for the PJR-PJN units in the one to three PPh units range, and the results were valid for all speech genres ($h=1, p<0.05$), except for WB.

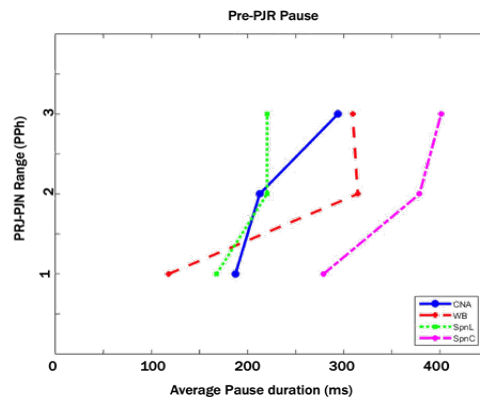


Figure 5. Average pre-PJR pause duration in correlation with projection size

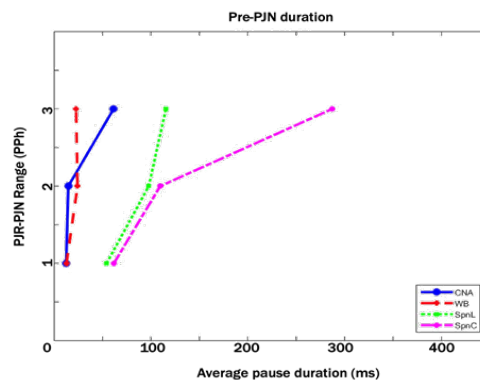


Figure 6. Average pre-PJN pause duration in correlation with projection size

4.3.2 Discussion

Based on the findings, it was suggested that when planning for a PJR-PJN unit as an information content unit, the speakers were mostly oriented to a longer pause in order to initiate the prosodic

highlights-prompted PJR unit. This further alluded to a longer preparation time required to plan for a longer projection trajectory. Although there were cases when a PJR-PJN unit with three PPh units was preceded by a slightly shorter pause, the general tendency mostly held given the statistical results of the pause durations for the PJR-PJN units in the one to three PPh units range. Another possible explanation was related to specific speech genre features. For the pre-PJR pause duration, the statistical results pointed to the main differences between read and spontaneous speech. We surmised that this reflected a discrepancy in the design of the speaking tasks in the four different speech genres: in the production of read speech, the speakers were given enough time to prepare before the actual recording; hence they had a chance to preplan the acoustic realizations for information projection due to familiarity with the reading materials. On the other hand, in the spontaneous speech genres, the planning of prosodic deployment and information content was interrupted intermittently because the spontaneous action that was interaction-based.

4.4 Correlation between Emphasis-attributed Weighting Scores and Information Projection

In the third analysis, we examined the correlation between emphasis-attributed weighting scores and information content projection. Following the findings reported in Chen and Tseng (2021) concerning the calculation of emphasis-attributed density scores throughout the PJR-PJN units, we further validated the information content loading distributions by prosodic highlights-prompted PJR-PJN units. It was demonstrated previously that speakers devote maximal efforts to the planning of information content from the beginning of prosodic highlights-prompted PJR units, and such effort decreases gradually throughout the projection trajectory (Chen & Tseng, 2021). However, Chen and Tseng (2021) reported the results of the emphasis-attributed weighting scores only by PJR-PJN units in the one PPh unit range. To extend the claim further, we carried out the weighting scores calculation again and included all the PJR-PJN tokens with a similar rationale and methodology proposed in Chen and Tseng (2021). We then conducted additional analysis of the correlation between the weighting scores and the PJR-PJN units in the one to three PPh units range for a more solid verification.

4.4.1 Results

As summarized in Figure 7, further analyses confirmed that, when a PJR-PJN unit was extended by three PPh units, a lower average emphasis-attributed weighting score was arrived at by the ending of the PJN units. In other words, the general trend of a decreasing weighting score following an increase in projection size was confirmed. This finding was quite consistent across the data of the four speech genres. Most of all, further *t*-test results verified that the average weighting scores were distinguished between the PJR-PJN units with one PPh unit and three

Prosodic Highlights-prompted Information Content Projection in Continuous Speech

PPh units. The statistical results were in general supported (all $h=1$, $p<0.05$), except for the spontaneous speech data from the SpnC genre.

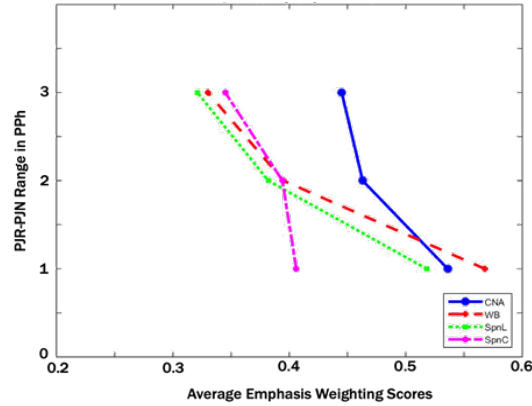


Figure 7. Correlation analysis between average emphasis-attributed weighting scores and the PJR-PJN units of the four speech genres (score assignment: $E_0=E_1=0$; $E_2=1$; $E_3=2$)

4.4.2 Discussion

Again, the above result confirmed that, when planning for prosodic-prompted projection for information content allocation, the speakers were oriented to a general pattern of heavy-to-light information loading across the PJR-PJN units, regardless of the projection trajectory size. When planning for a projection with a longer trajectory, the weighting scores decreased gradually toward the end of the projection, and hence information content loading diminished. Such findings in turn provided further confirmation of a PJR-PJN unit as the planning unit of prosodic highlights-correlated information content allocation and deployment in continuous speech. However, the statistical analysis did not find significant results for the SpnC data, which led us to wonder whether this may have had to do with the additional emphasis level of reduction (E_0) annotated for the current spontaneous speech genres (i.e., SpnL and SpnC). We attempted a further test by re-assigning the weighting scores only for the spontaneous speech data. In particular, we assigned a score of -1 to the emphasis level of reduction (E_0) annotated in the SpnL and SpnC genres, and then recalculated the average weighting scores. The results are summarized in Figure 8:

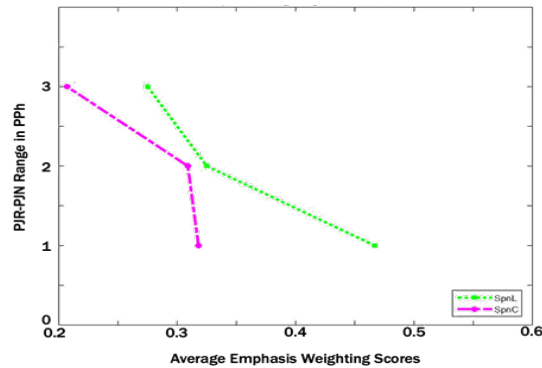


Figure 8. Correlation analysis between the PJR-PJN units and average emphasis-attributed weighting scores of the spontaneous speech genres (score assignment: $E0=-1$; $E1=0$; $E2=1$; $E3=2$)

Figure 8 presents a pattern similar to the above findings in that, the longer the PJR-PJN unit (i.e., up to three PPhs), the lower the average weighting scores derived from the ending of the PJN units. Further *t*-tests confirmed that the average weighting scores were distinguished between the PJR-PJN units with one PPh unit and three PPh units (both $h=1$, $p<0.05$). In other words, by taking into consideration the reduction annotation in the spontaneous speech genres, the heavy-to-light information allocation further stood out. With the attempt to faithfully model distinctive emphasis degrees in spontaneous speech signals, therefore, we were able to obtain even more solid evidence to support the current hypothesis regarding information content planning and allocation. This in turn verified the prosodic-prompted projections in association with information content deployment in continuous discourse and speech; above all, it was patterned on the prosodic highlights allotment in the speech context.

5. General Discussion and Summary

The current study focused on information content deployment that was prompted and projected by perceived prosodic highlights consistently annotated in continuous speech and discourse. In the first part of the analyses, we concentrated on the acoustic profiles of the information content planning of the PJR-PJN units, which was initiated and prompted by annotated tokens of prominence across four diverse speech genres. In terms of F0 realization, although the projection size differed in each PJR-PJN unit, we were able to derive a general falling contour starting with the PJR unit and throughout the whole projection trajectory. The further removal of higher-level intonation effects and the calculation of down-stepping degrees offered solid substantiations of the underlying intonation pattern. Above all, the current results demonstrated that only when we considered the information content planning unit of a global projection could we arrive at an identifiable falling contour with a clear down-stepping degree presented. Though the falling contour was within expectations and the results here are much in accordance with

Prosodic Highlights-prompted Information Content Projection in Continuous Speech

previous findings on prosody-based discourse units (i.e., Swerts & Geluykens, 1994), it was most crucial that we were able to further confirm that information content planning associated with prosody-prompted projections could possibly be established as a constant linguistic category with its own identifiable prosodic manifestation.

The second acoustic feature that we turned to was pause duration. As suggested, the duration of silent pauses located prior to the initiation of PJR and PJN units rendered some ideas about the relevant effort devoted to the planning of information content projection. It was demonstrated that, in order to plan for a longer projection, the speakers in general took more time prior to the initiation of the PJR and PJN units. Although not all pause-correlated results were presented with statistical significance, we assumed that the discrepancy was related to the task-specific features of the four difference speech genres.

Through the calculation of emphasis-attributed weighting scores, the third part of the analyses provided further validation of the “high-to-low” distribution of weighting score across the PJR-PJN units, which was similar to the finding from Chen and Tseng (2021). As previously indicated, the tendency of a higher weighting score for the initiation of information projection and a lower score for the end of information content projection reinforced the finding that the heaviest information loading was planned by prominence-prompted PJR units, with a gradually decreased planning effort demonstrated (Chen & Tseng, 2021). Here via the systematic modeling of prosodic highlights, including the reduction, our results faithfully reflected information content allocation and deployment for speech planning. Above all, the results showcased that only when taking into consideration the reduction feature in spontaneous speech could we arrive at a more significant distinction among the four speech genres with diverse features.

In sum, in this study we examined prosodic highlights-prompted information content planning and projection by the recently identified PJR-PJN units in continuous speech. Solid accounts were provided for the specific acoustic features, including F0 and pause duration, as well as the information-attributed weighting scores in correlation with the projection size in the PJR-PJN units. As has been identified previously the PJR-PJN units for information content projection were planned at a higher discourse-prosodic level from the HPG framework (cf. Chen & Tseng 2021); ultimately the identification of the patterns enabled a better understanding of information content planning within the hierarchical framework of the prosody context. In future studies, we propose to explore the following: (i) other possible acoustic correlates that might be involved prominence-prompted information content projection; (ii) empirical validations of the correlation between perceived prosodic distinctiveness in limited degrees and information weighting (i.e., Kurumada *et al.*, 2014); and (iii) the incorporation of the current analyses’ results into the automatic modeling of discourse prosody based on a hierarchical relationship (i.e., Lin *et al.*, 2019).

Acknowledgments

The authors gratefully acknowledge the assistance of Mr. Yen-Hsing Chen, Mr. Wei-Te Fang, and Dr. Chao-yu Su for the data analyses and relevant discussions in this study. Moreover, we would like to thank the research assistants from the Phonetics Lab at the Linguistic Institute, Academia Sinica (2015-2018) for the annotation tasks reported in this work. Most of all, we are grateful for the anonymous reviewers' and the editor's insightful comments on a previous version of the manuscript. All remaining errors, nevertheless, are ours.

About the Speech Corpora

To access the speech corpora incorporated in this study, the Sinica COSPRO corpus (for the read speech data) can be obtained through the database link from the Association for Computational Linguistics and Chinese Language Processing (http://www.aclclp.org.tw/use_mat.php#cospro). A sample of the spontaneous interaction (SpnC) data with the annotations described in this paper can be obtained via contacting the first author.

Abbreviations

2SG	second-person singular pronoun	COP	copula
1PL	first-person plural pronoun	DE	associative/complementizer de
CL	classifier		

References

- Auer, P. (2015). The temporality of language in interaction: Projection and latency. In A. Deppermann & S. Günthner (Eds.), *Temporality in interaction*, 27-56. John Benjamins.
- Baumann, S., Niebuhr, O., & Schroeter, B. (2016). Acoustic cues to perceived prominence levels: Evidence from German spontaneous speech. In *Proceedings of 8th Speech Prosody Conference*, 711-715. <https://doi.org/10.21437/SpeechProsody.2016-146>
- Boersma, P., & Weenink, D. (2015). Praat: Doing phonetics by computer. Retrieved November 20, 2015, from <http://www.praat.org>
- Chen, H. K., Fang, W., & Tseng, C. (2016). Advance prosodic indexing-Acoustic realization of prompted information projection in continuous speeches and discourses. In *Proceedings of the 10th International Symposium on Chinese Spoken Language Processing (ISCSLP 2016)*. <https://doi.org/10.1109/ISCSLP.2016.7918412>
- Chen, H. K., Prévot, L., Bertrand, R., Priego-Valverde, B., & Blache, P. (2012). Toward a Mandarin-French corpus of interactional data. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogues*, 147-148.

Prosodic Highlights-prompted Information Content Projection in Continuous Speech

- Chen, H. K., & Tseng, C. (2021). From speech to language—An alternative corpus account of prosodic highlight in continuous speech. *Concentric: Studies in Linguistics*, 47(2), 184-224. <https://doi.org/10.1075/consl.00027.che>
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181-204. <https://doi.org/10.1017/S0140525X12000477>
- Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, 40(2), 141-201. <https://doi.org/10.1177/002383099704000203>
- De Saussure, F. (1966). *Course in general linguistics*. McGraw-Hill Book Company.
- Dilley, L. (2016). Rhythm, context effects, and prediction. In *Proceedings of Speech Prosody 2016 conference (SP 2016)*.
- Falk, S. (2014). On the notion of salience in spoken discourse—Prominence cues shaping discourse structure and comprehension. *TIPA, Travaux Interdisciplinaires Sur La Parole et Le Langage [Interdisciplinary Works on Speech and Language]*, 30, 1-23. <https://doi.org/10.4000/tipa.1303>
- Goodwin, C. (1996). Transparent vision. In E. Ochs, E. A. Schegloff & S. A. Thompson (Eds) *Interaction and grammar*. Cambridge University Press. <https://doi.org/10.1017/cbo9780511620874.008>
- Halliday, M. A. K. (1967). Notes on transitivity and theme in English: Part 1. *Journal of Linguistics*, 3(1), 37-81. <https://doi.org/10.1017/S0022226700012949>
- Kohler, K. J. (1997). Modelling prosody in spontaneous speech. In S. Yoshinori, N. Campbell & N. Higuchi, (Eds.), *Computing prosody: Computational models for processing spontaneous speech*, 187-210. Springer. https://doi.org/10.1007/978-1-4612-2258-3_13
- Kurumada, C., Brown, M., Bibyk, S., Pontillo, D., & Tanenhaus, M. (2014). Rapid adaptation in online pragmatic interpretation of contrastive prosody. In *Proceedings of the Annual Meeting of the Cognitive Science Society 2014*, 791-796
- Lehiste, I. (1970). *Suprasegmentals*. MIT Press.
- Lieberman, P. (1967). *Intonation, perception, and language*. MIT Press.
- Lin, C. H., You, C. L., Chiang, C. Y., Wang, Y. R., & Chen, S. H. (2019). Hierarchical prosody modeling for Mandarin spontaneous speech. *Journal of Acoustic Society America*, 145(4), 2576–2596. <https://doi.org/10.1121/1.5099263>
- Pierrehumbert, J. B., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Morgan, J. L. Cohen, & M. E. Pollack (Ed.) *Intentions in communication*. MIT Press.
- Silverman, K., Beckman, M., Pitrelli, J. Ostendorf, J. M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. In *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP 1992)*, 867-870.

- Swerts, M., & Geluykens, R. (1994). Prosody as a marker of information flow in spoken discourse. *Language and Speech*, 37(1), 21-43. <https://doi.org/10.1177/002383099403700102>
- Tseng, C. (2010). An F0 analysis of discourse construction and global information in realized narrative prosody. *Language and Linguistics*, 11(2), 183-218.
- Tseng, C. (2013). Output prosody-How information highlights are piggybacked by discourse structure. *Chinese Journal of Phonetics (中國語音學報)*, 4, 109-124.
- Tseng, C., Cheng, Y., Lee, W., & Huang, F. (2003). Collecting Mandarin speech databases for prosody investigation. In *Proceedings of the 2003 International Conference Oriental-COCCOSDA*.
- Tseng, C., Lee, L., & Su, Z. (2008). Spontaneous Mandarin speech prosody - The NTU DSP lecture corpus. In *Proceedings of the 2008 International Conference Oriental-COCCOSDA*, 171-174.
- Tseng, C., Pin, S., Lee, Y., Wang, H. & Chen, C. (2005). Fluent speech prosody: Framework and modeling. *Speech Communication*, 46(3-4), 284-309. <https://doi.org/10.1016/j.specom.2005.03.015>
- Tseng, C., Su, C. & Huang, C. (2011). Prosodic Highlights in Mandarin Continuous Speech—Cross-Genre Attributes and Implications. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, 1381-1384. <https://doi.org/10.21437/Interspeech.2011-454>
- Tseng, C., & Su, Z. (2008). Discourse prosody and context—global F0 and tempo modulations. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008)*, 1200-1203. <https://doi.org/10.21437/Interspeech.2008-361>
- Tseng, C., & Su, Z. (2012). Information allocation and prosodic expressiveness in continuous speech: A Mandarin cross-genre analysis. In *Proceedings of the 2012 International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 243-246. <https://doi.org/10.1109/ISCSLP.2012.6423535>
- Watson, D. G., Tanenhaus, M. K., & Gunlogson, C. A. (2008). Interpreting pitch accents in online comprehension: H* vs. L + H. *Cognitive Science*, 32(7), 1232-1244. <https://doi.org/10.1080/03640210802138755>

Topic Development and Boundary Cues in Hakka Conversational Discourse

Shu-Chuan Tseng* and Hsiao-chien Liu⁺

Abstract

The structure of conversational discourse is context-dependent, and the organization of discourse segments and preferences for signaling discourse boundaries are language-specific characteristics. Participating speakers, speaking scenarios, and communication purposes instantaneously affect the conduct of social interaction and verbal exchanges during a conversation. For example, topic maintenance is sustained by the overt exchange of coherent information, and lexical preferences at the boundaries of related discourse segmentation can help construct the course of topic development. Moreover, form-based discourse units are used to represent the content of spoken utterances and to describe the interaction of speakers in conversations. This study investigated topic-specific Hakka conversations using a top-down two-level discourse segmentation approach to examine the development and production of topics. Typical cues and expressions used to initiate topics and subtopics and their respective discourse functions in the Hakka conversations were analyzed. In the Hakka conversational data, noun phrases were preferred at the topic and subtopic transition boundaries, and complete forms such as clausal constructions were also favored, although the spontaneous speech was expected to be fragmentary in terms of syntactic structure.

Keywords: Conversation, Discourse Units, Topic Development, Boundary Cues, Hakka

1. Introduction

A constituent of a given discourse may be defined as a “contextually indexed representation of information conveyed by a semiotic gesture, asserting a single state of affairs or partial state of affairs in a discourse context,” as proposed by Polanyi (2005: 266). This kind of discourse

* Institute of Linguistics, Academia Sinica

E-mail: tsengsc@gate.sinica.edu.tw

⁺ College of Hakka Studies, National Central University

E-mail: justcarrie1@gmail.com

segment involves interactive domains, such as discourse genres and speech events, and its segmentation is mainly guided by semantic criteria (e.g., a complete state of affairs and a complete semantic representation), syntax (e.g., clauses and sentence boundaries), and intonation (e.g., pauses and prosodic contours) (Polanyi, 1995). Moreover, discourse segments are indicated by topic shift markers that have been categorized as discourse markers, pragmatic markers, discourse operators, and cue phrases in the literature (van Dijk, 1977b; Grosz & Sidner, 1986; Fraser, 1996; Redeker, 2006). Polanyi (1995) also proposed that discourse operators force segmentation breaks on semantic grounds, as will be shown later in the data from the Hakka conversations. To describe the semantic structure of a conversational discourse, constituent units and their composition/decomposition principles are needed as well as devices to identify boundaries for effective discourse segmentation.

1.1 Discourse Topics

Discourse topics form a coherent discourse that expands on a number of common themes. Van Dijk (1977a: 136) defined discourse topics as “a proposition entailed by the joint set of propositions expressed by the sequence...proposition T is TOPIC of sequence of propositions $\Sigma = \langle P_1, P_2, \dots, P_n \rangle$ iff for each $P_i \in \Sigma$ there is a subsequence Σ_k there is a P_j such that $\Sigma_k \Rightarrow P_j$ and $T \Rightarrow P_j$ ”: each sequence entails a global proposition P_i , and the global proposition entails a super-global proposition P_j , which is the topic. Giora (1985: 116) also defined “the element relative to which the whole set of propositions (of that segment) is taken to be ‘about’”; in other words, a topic should not be derivable from the discourse that occurs before it is introduced. Slightly differently, Geluykens (1993: 118) defined a topic as “information which has a low degree of Recoverability...and which has Persistence”: for information to be considered a topic, it must be sustained over a reasonably long stretch of discourse. Stede (2012: 38) gave a clear definition: “A topic as a property of a text segment is characterized by the particular distribution of content words in that segment, and the difference to the distribution in other segments.” Todd (2016: 11) combined his own definition with Giora’s (1985) “aboutness” and proposed that a topic is determined on the basis of aboutness, connectedness, and relevance. Connectedness is relevant to cohesion and coherence and makes a stretch of language into a meaningful whole. Topics are usually distinguished in terms of their explicitness, with cohesion used for explicit links, or the overt relationship between propositions, and coherence for implicit links, which requires background knowledge or contextual knowledge for interpretation. To identify topic boundaries, cohesion markers, lexis, and coherent concepts are applied. Aboutness is a semantic construct in which all propositions in the discourse are related to a superordinate discourse topic; relevance is concerned with the relationship between a proposition and the one that precedes it, and consistent relevance between propositions makes a discourse coherent.

Asher (2004) applied formal semantic analysis and developed a dynamic theory of

discourse topics by examining four types of contrastive topics: (1) alternation, which incorporates parallel and contrastive notions; (2) narration, which represents two connected discourse segments that appear in a background-foreground event; (3) subordinating and coordinating relationships, which are dependent on the degree of attachment to the antecedents; and (4) summarizers, which are used when there are many discourse segments. Furthermore, Asher and Vieu (2005) noted that in the segmented discourse representation theory, a common topic can be shared by two related constituents (Asher, 1993). Regarding the principles within a topic, the “right frontier constraint” provides attachment points for new information (Webber, 1988: 8). Another proposed principle—continuing discourse patterns—suggested that a coordinating relationship bears the same discourse relationship with a dominating constituent and that the coordinated constituents of a substructure must follow a certain pattern with respect to the dominating constituent (Asher & Vieu, 2005: 595). Subordination and coordination affect topicality in that two constituents are coordinately linked if they contribute to the topic of the larger segment, while they are subordinately linked if one of them is a subtopic (Asher, 1993; van Kuppevelt, 1995a).

Givón (1983) proposed a hierarchical structure that accounted for the preceding discourse context information. In macro structures, thematic paragraphs are larger thematic units that are composed of multi-propositional and chained clauses. Within a thematic paragraph, there are three types of topics—chain initial topics, chain medial topics, and chain final topics—defined by their relative position in a speech flow. A chain initial topic is a “newly introduced, newly changed or newly returned topic” (Givón 1983: 9), and therefore usually has a discontinuous relationship with the preceding context but is potentially persistent in the succeeding context if it introduces an important topic. A chain medial topic is continuous in terms of the preceding context and is persistent, but not maximally so, in the succeeding context. Finally, a chain final topic is continuous in terms of the preceding context but is not persistent in the succeeding context, even if it deals with an important topic. Givón (1983) also defined three quantitative measures—referential distance (“lookback”), potential interference (“ambiguity”), and persistence (“decay”)—to describe topic properties in discourse; these measures reflect the degree of topic continuity, topic disruption, and topic persistence, respectively.

Van Kuppevelt (1995a, 1995b) proposed that the topic unit does not always stick to the NEW/OLD principle but appears in different syntactic forms. He further specified that

the main structure of a bound discourse is determined by one leading discourse topic constituted in one production step at the beginning of the discourse. The development of such a discourse is, with regard to its main structure, from the beginning, bound programmatically by the set topic-constituting questions defining its discourse topic. (van Kuppevelt, 1995a: 139)

The topic hierarchy of a discourse, according to this proposal, contains discourse topics, topics, and subtopics.

1.2 Discourse Segmentation

Discourse segments can be of various sizes and empirically specified by applying operational principles that define their linguistic forms and discourse functions. Within a discourse topic, there is normally a kind of coherent relationship between adjacent lexical chains, which reflects the meaning and function of the discourse. In a conversation, a degree of unity is required to achieve cohesive relationships between sequences of words within a certain stretch of speech, such as reference, ellipsis, substitution, conjunction, and lexical cohesion (Halliday & Hasan, 1976). Coherent relationships between clauses and sentences, such as elaboration, subordination, cause, and exemplification, were also discussed in depth by Mann and Thompson (1988). Morris and Hirst (1991) analyzed lexical cohesion to determine coherence by computing lexical chains in a thesaurus, where thesaural relationships, transitivity of word relations, and distance in sentences allowable between words in a chain were examined. Hoey (2005) used the convergence of overt cohesion and perceptible coherence as the criteria and found that lexical priming and cohesion influenced the comprehensibility of a discourse's organization.

When a topic chain occurs over a succession of several nearby clauses that share a single topic, topic shifts completely direct the discourse text away from the present topic, while topic drifts do not stray far from the present topic. A topic returns if it is mentioned again. Based on Hobbs (1990), three coherence relationships regarding topic drifts have been proposed: if two segments assert propositions with similar or identical properties, then they have a parallel relationship; if a segment serves as a cause for another segment, then this represents an explanation relationship; and if a segment involves the evaluation of comments on a previous topic with no additional new information, then it has an evaluation, or metatalk, relationship.

Cues in topic shifts indicate digressions, but cues in topic drifts may not. Considered in light of Fraser's (1996, 2009) definitions, topic shift cues correspond to topic change markers or digression markers, while topic drift cues can link to other markers of different functions, such as contrast, elaboration, and inference. Todd (2016) considered that there is a continuum from shift and drift to maintenance, rather than these cues forming discrete categories of boundaries. More specifically, drifts are weak boundaries, whereas shifts are strong boundaries. In English, topic boundaries may be marked to signal a shift and attract attention, as in the case of 'Oh, I meant to tell you'. Conversely, 'well' is likely a topic drift marker since it has a much more ambiguous role and is followed by a mix of new and old information. Topic shift markers are used to indicate discourse boundaries, but because of different perspectives and a long history of investigations, they have been given various labels, such as pragmatic connectives (van Dijk, 1977c), discourse particles (Schourup, 1985), discourse markers, transition markers,

discourse operators (Schiffrin, 1987; Fraser, 2006; Redeker, 2006), pragmatic markers (Fraser, 2009), digressive markers (Charolles, 2020), cue phrases (Grosz & Sidner, 1986; Hirschberg & Litman, 1993; Horne *et al.*, 2001), clue words (Cohen, 1984), and so on. Quirk (1972) proposed a system of taxonomy that included parallel, inference, summary, detail, reformulation, and contrast markers illustrated by the cue phrases ‘in addition’, ‘as a result’, ‘in sum’, ‘in particular’, ‘in other words’, and ‘conversely’, respectively. Grosz and Sidner (1986), Fraser (1996), and

Table 1. Three proposals for boundary cues

	Functions	Examples
Cue phrases	attentional change	(push) now, next, that reminds me, and, but (pop to) anyway, but anyway, in any case, now back to (complete) the end, ok, fine, (paragraph break)
	true interruption	I must interrupt, excuse me
	flashback	Oops, I forgot
	Grosz & Sidner (1986: 198)	digression
	satisfaction-precedes	in the first place, first, second, finally, moreover, furthermore
	new dominance	for example, to wit, first, second, and, moreover, furthermore, therefore, finally
Turn-internal discourse segment transitions in spontaneous speech	end of segment	okay?, you know, so
	next segment	okay, so, but, now, well, and
	digression, interruption	by the way, you know
	specification, definition	that is, you know, well
	paraphrase	I mean, you know, that is
	explication, clarification	because, you know, I mean
	background information	because, see, well
Redeker (2006: 345)	comment	you know, I think, I guess
	correction, emendation	oh, or, I mean
	quote	you know, like, well, oh
	return	but (anyway), so, now, well
Pragmatic markers	topic change markers	incidentally, speaking of X, parenthetically, by the way, just to update you, that reminds me, before I forget, back to my original point, returning to my point, on a different note
	contrastive markers (denial or contrast)	but, instead, however, all the same, anyway, in any case/rate/event, nevertheless, conversely, despite, even so, regardless, still, that said, though, yet
Fraser (1996)	elaborative markers (elaboration or refinement)	above all, in other words, what’s more, also, alternatively, besides, by the same token, correspondingly, for instance, on top of it all, to cap it all off
	inferential markers (developed based on inference)	after all, so, accordingly, because of this/that, for this/that reason, it can be concluded that, it stands to reason that, of course, then, thus, so

Redeker (2006) all intended to capture the indications of segment transitions, but Fraser's (1996) four discourse markers are much more straightforward. The three proposals for boundary cues are summarized in Table 1.

Boundary cues, as defined in terms of the listed discourse functions in Table 1, are of various linguistic lengths. Words, word sequences, short phrases, and clauses can all serve as boundary cues. In addition to the issue of linguistic units, the conventional use of discourse markers may not correspond precisely to the function of marking topic boundaries because discourse markers are defined as expressing distinctive functions, not as indicating coherent relationships between discourse segments (Harabagiu, 1999). According to the data presented by Das (2014), a majority of topic shift relationships are not explicitly signaled by discourse markers.

This study aimed to investigate the discourse structure of Hakka conversations by applying a top-down two-level annotation schema of discourse segments that form sequences of lexical chains with coherent relationships within sequences and cohesive relationships across sequences. The continuity and maintenance of coherent and cohesive relationships were used as the main judgment criteria for identifying the boundaries of discourse segments. Furthermore, we used form-based units to represent the linguistic content to describe the local environment of the lexical chains. The words and phrases that occurred at the topic and subtopic boundaries were then investigated in the context of discourse segmentation.

2. Method

2.1 Data

This study examined five Taiwan Hakka conversations recorded for the National Digital Language Archive Project. The conversations were produced by six female and four male native Hakka speakers aged between 34 and 60 years old. There are two major variants of Taiwan Hakka: Hailu and Sixian (Hakka Affairs Council, 2017). Our sample included five Hailu and five Sixian speakers. All 10 speakers reported that they were fluent in Hakka and that they spoke Hakka better than Taiwan Mandarin and Taiwanese Southern Min. The pairs of speakers were instructed to talk about a topic of their choice, and each recording session lasted approximately 15 minutes. The content of the conversations was lexically transcribed by the second author whose mother tongue is Sixian Hakka. Word segmentation and part-of-speech tagging were conducted according to the *Dictionary of Frequently Used Taiwan Hakka* published by the Ministry of Education in Taiwan. However, some cases did require discussions and consultations with native speakers and linguists. For example, the negation 無 *mo55* in 美術先生無教个 *mui31 sud2 sin31 sang31 mo55 gau31 gai11* ('what the art teacher did not teach') in one of the Hailu Hakka conversations was listed as an auxiliary word in the dictionary, but

in authentic usage, it can also be a verb, a negative marker, or a negator, depending on the context.

2.2 Annotation of Topics and Subtopics

When the main focus of attention shared by the conversational partners changes, it is considered a topic shift. A conversation segment is labeled “topic” if the concepts and messages exchanged by the interlocutors form a coherent set of interactions. A topic segment in principle presents a high degree of coherence and cohesion in terms of the connectedness, relevance, and aboutness of the information expressed in the conversation. A topic segment can only be annotated if the entire stretch shows a steady and continuing context with topic maintenance. In the framework of lexical cohesion analysis, a straightforward way to identify topic boundaries is to focus on lexical chains (Todd, 2016: 41-43), particularly content word collocations or conceptual associations such as boys-girls, laugh-joke, and bee-honey (Halliday & Hasan, 1976; Todd, 2016). Within each topic, there can be a series of components of interactions that represent different manners of elaborating the topic, and these components are annotated as “subtopics.” Subtopics normally appear sequentially but may also overlap and recur, as responses from conversational partners are spontaneous. The identification of subtopic boundaries relies on crucial phrases that establish the relationship of lexical cohesion and that serve as the main clues. For instance, for the topic “family,” subtopics such as “places of residence” and “children” may be annotated by the names of places and family members or jobs that are reiterated in consecutive utterances. Depending on the research questions and approaches, there may be different segmentation schemes of conversational discourse, and the annotation of discourse structures is to some degree subjective.

In the current study, we used the two-level discourse segmentation scheme presented above and implemented a procedure to possibly mitigate the level of subjectivity. The five Hakka conversations were first transcribed by a native Hakka speaker and then translated into Mandarin texts that were proofread by three adult native Mandarin speakers. Segmentation into topics and subtopics was conducted by the authors by applying the above principles. Two independent annotators were recruited to evaluate whether the identified boundaries of the topics and subtopics were appropriate for segmenting the conversations. Table 2 lists the results. Both annotators reached agreement in nearly 80% of the topic and subtopic boundaries assigned by the authors. We noticed that in one of the conversations, the rate of disagreement was particularly high, which may have been attributed to a large number of unclear transitions held by the very dominant speaker who produced long topic segments that consisted of several subtopics without a clear boundary. At least one annotator or both annotators did not agree with 20% of the originally segmented boundaries. The location of the boundary was generally agreed by both annotators. Disagreements mostly resulted from deviated judgement about whether a

boundary was a subtopic or a topic. These boundaries were reconsidered and revised by the authors. Eventually, we obtained a final version of discourse segmentation for the Hakka conversations.

Table 2. Topic and subtopic boundary labeling

Boundaries	Hakka Conversations
# of topic boundaries	46
# (%) in agreement	36 (78.26%)
# of subtopic boundaries	261
# (%) in agreement	214 (81.99%)

Below is an excerpt from the data showing a discussion on the topic “language use.” The subjects 吾家娘 (‘my mother-in-law’) and 佢个細人仔 (‘my child’) were often omitted in the utterances. However, this nominal ellipsis did not hinder the participants’ understanding of the speech content, and the topic “language use” clearly remained the focus of successive elaborations until a conclusion was finalized at the end of this discourse segment.

吾家娘乜當希望佢个細人全部講客話	<i>nga55 ga31 ngiong55 me11 dong53 hi53 mong33 ngai55 gai11 sel1 ngin55 cion55 pu33 gong24 hag5 fa11</i> (‘ <u>my mother-in-law</u> also expected <u>my child</u> to speak Hakka all the time’)
佢渡个時節	<i>gi55 tu33 gai11 shi55 zied5</i> (‘when <u>she [my mother-in-law]</u> took care of <u>[my child]</u> ’)
全部講客話 hon	<i>cion55 pu33 gong24 hag5 fa11 hon</i> (‘ <u>[my mother-in-law]</u> spoke Hakka all the time’)
可是讀書開始	<i>ko24 shi33 tug2 shu31 koi53 shi24</i> (‘but since <u>[my child]</u> started going to school’)
讀幼稚園開始	<i>tug2 rhiu11 chi55 rhan55 koi53 shi24</i> (‘since <u>[my child]</u> started going to kindergarten’)
斯專門講國語啊	<i>sii53 zhon53 mun55 gong24 gued5 ngi53 a</i> (‘ <u>[my child]</u> just spoke Mandarin day and night’)
佢成時講國語	<i>gi55 shang55 shi55 gong24 gued2 ngi31</i> (‘ <u>she [my child]</u> spoke Mandarin constantly’)
啊國語講啊流流利利	<i>a gued5 ngi53 gong24 a liu55 liu55 lad3 lad3</i> (‘ <u>[my child]</u> spoke Mandarin fluently’)
講久	<i>gong24 giu24</i> (‘ <u>[my child]</u> had been speaking for a long time’)
該客話斯毋記得了 hon	<i>gai55 hag5 fa11 sii33 m55 gi11 ded5 le31 hon</i> (‘ <u>[my child]</u> did not know how to speak Hakka’)

2.3 Annotation of Discourse Units

We used a form-based discourse unit (DU) to represent and analyze the discourse structure of the sampled conversations concerning the more information-based discourse segments, topics, and subtopics. In principle, a DU is equivalent to a clause or a sentence in written language. After the main predicate is identified, a DU includes the speech stretch containing the main predicate and the remaining syntactic components, including the subject and the related complements and adjuncts. Some DUs are isolated noun phrases or non-clausal units with no predicates. Non-predicative DUs of this type occur frequently in Japanese and Mandarin Chinese interactional discourse and employ a range of functions, such as referent introduction, identification, and listing (Iwasaki, 1993; Tao, 1996, 2020). In Hakka, verb complexes are often used. To identify DUs in the Hakka data, we referred to the definition of clauses proposed by Thompson and Couper-Kuhlen (2005) and the principles of determining utterance units (Nakajima & Allen, 1993). Please note that DUs are annotated solely based on their constructional form and that both predicative and non-predicative DUs can express complete or incomplete meanings and information. We proposed dividing the DUs into three main types according to their form and meaning: (i) clausal DUs with clear meaning; (ii) non-clausal DUs with clear meaning; and (iii) fragmentary DUs with incomplete meaning. The linguistic content of the Hakka conversations was represented and analyzed in terms of DUs and DU types. Detailed explanations of the DU annotation principles are provided below:

(1) Clausal DUs with clear meaning

- a. Clauses delineate complete sentential meanings and satisfy discourse functions, e.g., 佢就渡一個細嬰 *gi55 ciu33 tu33 rhid5 gai11 se11 o53* ('He only takes care of one baby') and 佢會講分佢俗仔聽喔 *ngai55 voi33 gong24 bun53 ngai55 lai11 er55 tang11 o* ('I will tell my son!'). These kinds of DUs often express substantial statements in conversations.
- b. DUs with elliptical subject or object NPs that convey a clear and coherent meaning, e.g., 敢還哪看得著客家話 *gam31 han11 nai55 kon55 ded2 do31 hag2 ga24 fa55* ('Where can we see Hakka language?!'), 聽毋識 *tiang24 m11 siid2* ('[I] cannot understand'), 來正知个啊 *loi55 zang11 di31 ge55 a* ('Only when we came did they know that [we are Hakka]'), and 面前就講 *mien55 qien11 qiu55 gong31* ('[I] talked about it earlier').
- c. Complex DUs that contain focus markers¹ or conditional markers, e.g., 無講若般看人斯做毋得 *mo55 gong24 rhog2 ban53 kon11 ngin55 sii53 zo11 m55 ded5* ('Not to

¹ The typical Mandarin Chinese focus marker in the cleft constructions 是 *shi* and 是...的 *shi...de* are not exactly the same as 無講, 斯, 係 in Hakka. In this study, they were tentatively considered the focus markers that served the function of emphasis or indications of the upcoming discourse segment.

mention it is not permitted to have a short look at people'), 恁自家愛想愛去哪斯去哪 *an31 qid2 ga24 oi55 xiong31 oi55 hi55 nai55 sii24 hi55 nai55* ('So I go out at will'), and 係無貪就無熟事 *he55 mo11 tam24 qiu11 mo11 sug5 sii55* ('If [you] are not greedy, you will not know [those people]').

(2) Non-clausal DUs with clear meaning

- a. This type of DU has no predicate but conveys a clear discourse meaning. These DUs may be used for responses or to introduce a new topic and can take a variety of constructional forms, e.g., 係啊 *he55 a* ('yes'), 正經啊 *ziin55 gin24 a* ('[It is] real'), 吾嫂這兜啊 *nga24 so31 ia31 deu24 a* ('[this situation] applies to people like my sister-in-law'), and 僱苗栗縣个 *ngai11 meu11 lid5 ien55 ge55* ('I [am from] Miaoli County').²
- b. Predicative adjectives used as part of a verbal complement, e.g., 若若若細人幾大 *ngia24 ngia24 ngia24 se55 ngin11 gi31 tai55* ('how old are your children'), 補助盡高喔 *bu31 cu55 qin55 go24 o* ('the subsidy is high!'), and 恁打爽忒了 *an24 da24 song24 ted5 le53* ('that is a pity'). Thompson and Tao (2010) found that conversational Mandarin speech favors predicate adjectives (80%) over attributive adjectives (20%).
- c. Particle DUs that are responsive backchannels, such as 嗯 *n*, 嗯嗯 *nn*, 喔喔 *oo*, and 唉 *ai*, or connective-like junctures that express different speaker attitudes. For instance, the modal particle *hon* in the following example serves as the concluding function and expresses the speaker's intention to obtain approval from the conversational partner in *hon*, e.g., *hon...<我們在畫畫>个時節佢會攞人<修改> hon...wo214 men zai51 hua51 hua51 gai11 shi55 zied5, gi55 voi33 lau31 ngin55 xiu55 gai214* ('hon...he would help students make modifications when we were drawing').

(3) Fragmentary DUs with incomplete meaning

- a. DUs that contain noun phrases are used to express the speaker's intention or for communicative functions, such as introducing referents (Iwasaki, 1993; Tao, 1996, 2020). Isolated noun phrases are seldom used to indicate topic shifts and drifts. They may appear together with prepositions or particles, e.g., 對厥印象 *dui11 gia55 rhin11 siong33* ('about the impression of him') and 然後在<那個>³ 年代 *hon 恁仔 rhan55 heu33 cai33 na31 ge ngien55 toi33 hon an24 ne31* ('then, in that era').
- b. Disfluent DUs with no predicates, such as speech repairs or repetitions, e.g., 僱一句僱

² 僱苗栗縣个 is considered a non-predicative DU with the nominalization marker 个.

³ Content appearing in <> was spoken in Mandarin Chinese.

又毋 *ngai11 id2 gi55 ngai11 iu55 m11* ('I cannot even [say] a sentence...') and 係毋識 *ngai55 m11 siid2* ('I have not been...'). Please note that repairs are not necessarily related to the proposition of the next DU uttered by the same speaker or by the conversational partner, e.g., 噃恁仔關於講該教育方面个時節...以前个時節你會 (repair)...啊比論講啊你以前無讀著个理想个 *n an24 e31 gon31 rhi55 gong24 gai55 gau11 rhug2 fong31 mien11 gai11 shi55 zied5...rhi31 cien55 gai11 shi55 zied5 ngi55 voi33...a bi24 lun33 gong24 a ngi55 rhi31 cien55 mo55 tug2 do24 gai11 li31 siong24 gai11* ('As for education...before, you would...for example, you have not majored in the ideal subjects...').

2.4 Results

The annotation results of the topics, subtopics, and DUs in the five Hakka conversations are summarized in Table 3. Each conversation covered a different number of distinctive topics that were initiated and discussed by the participants. Please note that the speakers may have restarted a previously discussed topic initiated by themselves or their conversational partners along the course of the conversation. In such cases, we included the occurrences of returning topics in the calculation of topic segment tokens. The number of subtopics per topic was between four and five, but the patterns of speaker interaction and information exchange were in fact individually different in the Hakka conversational data, which will be shown later.

Table 3. Annotation results of the five Hakka conversations

	Con. 1	Con. 2	Con. 3	Con. 4	Con. 5
Duration	11 mins	11 mins	14 mins	13 mins	11 mins
# of topics	9	8	11	10	8
# of subtopics	45	53	40	56	31
# of topic segments	12	10	12	13	8
# of subtopic segments	53	54	47	67	35
# of DUs	665	598	878	715	554
# of syllables	3,377	3,320	3,687	3,626	2,810

While the complete coverage of concept exchanges was sustained within a topic, subtopics were operationally more authentic in that they actually formed the continuity of the topic, on the one hand, and connected the consecutive DUs, on the other hand. This also indicated that the ratio of DUs to subtopics was an authentic reflection of topic transitions. The referential distance measurement in Givón's (1983) hierarchical structure proposed that the degree of distancing in topic continuity is 20 clauses in terms of the number of clauses toward the left edge. Additional attempts to measure topic segment length have been proposed in the literature,

for instance, a typical paragraph (Ferret & Grau, 2000), a length of three to five sentences (Hearst, 1993), and a length of approximately 100 words (Dias & Alves, 2005). As shown in Table 3, the degree of speaker activity in the conversations varied, as the number of topics and subtopics initiated by each speaker was considerably different.

The complexity of topics and subtopics to some degree revealed idiolect differences in maintaining topic continuity. Nevertheless, collective commonalities across the Hakka speakers were shown by the number of syllables per DU, which ranged from four to six. Prévot *et al.* (2015) examined DU distributions in French and Taiwan Mandarin conversational data and reported an average DU length of 10.7 syllables for French and 9.6 syllables for Taiwan Mandarin in long speaker turns. Prévot *et al.*'s (2015) study mainly focused on DU components rather than speaker interaction. Compared with Givón's (1983) proposed measure of 20 clauses over a sustained topic, our measurement of DUs per subtopic in Table 4 showed similar results:

Table 4. Annotation results of the five Hakka conversations by speaker

	Con. 1		Con. 2		Con. 3		Con. 4		Con. 5	
Speaker gender	F	M	F	M	F1	F2	F	M	F	M
# of syllables	1,158	2,219	1,502	1,818	1,703	1,984	2,709	917	1,289	1,521
# of distinctive topics	5	4	3	5	3	8	7	3	1	7
# of distinctive subtopics	21	24	26	27	18	22	41	15	12	19
# of DUs	294	371	260	338	396	482	471	244	268	286
# of syllables per DU	3.9	6	5.8	5.4	4.3	4.1	5.8	3.8	4.8	5.3
# of topic initiations	6	6	4	6	3	9	10	3	1	7
# of subtopic initiations	26	27	26	28	21	26	52	15	14	21
# of DUs per topic segment	49	61.8	65	56.3	132	53.6	47.1	81.3	268	40.9
# of DUs per subtopic segment	11.3	13.7	10	12	18.9	18.5	9.1	16.3	19.1	13.6

In addition, more topic and subtopic initiations did not imply more DU production, as shown in Table 4. That is, the degree of active participation in the verbal exchanges was viewed from different perspectives. For instance, in conversation #4, the male speaker clearly initiated fewer new topics, but his active participation was supported by the large number of DUs he produced in taking part in the discussion. On the other hand, compared with his counterpart, he produced shorter DUs that did not deliver complex information as they were mostly responsive. In our two-level discourse segmentation approach, whether topic initiation occurred in response to the previous information delivered by the conversational partner was also an important clue in determining the degree of active participation.

3. Discourse Organization in Hakka Conversations

Understanding conversational discourse requires a structural description of how the discourse is organized. Therefore, it is necessary to have a system of segmentation units whose relationships can be empirically defined. In this study, we annotated three types of units—topics, subtopics, and DUs. Topics and subtopics were identified from a top-down perspective, in which the information content was the principal criterion. The DUs were mainly identified according to their constructional forms. In particular, predicates were used to categorize the types of DUs.

3.1 Conversational Discourse Descriptions

Todd (2016: 172) mentioned that topic development is divided into three main types—maintenance, drift, and shift—that can be further categorized into subtypes, such as major and minor shifts. In our annotation of topics and subtopics, we took these main types into consideration to describe the interplay of information exchanges and transactions in the discourse organization of the Hakka conversations. Figure 1 shows the hierarchical structure of the topics and subtopics represented by the DUs extracted from the data. The DUs produced by *Speaker A* are underlined in the transcript, and speech content uttered at topic and subtopic boundaries are in boldface. The interaction of the speakers and the speech production patterns is illustrated in terms of this representational format. The identification of topic shift and drift was content-based, while the DUs were defined according to the placement and scope of the predicates.

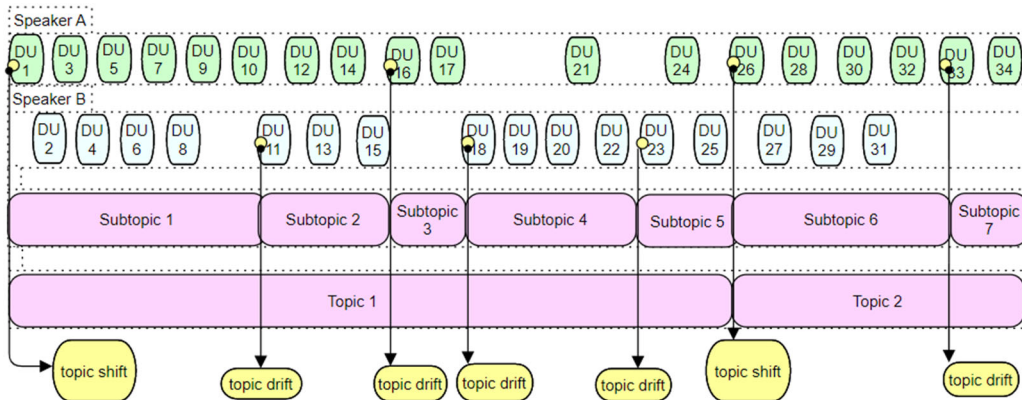


Figure 1. Topic development in the conversations

Topic 1: Speaking Hakka⁴

Subtopic 1: It's difficult for Southern Min people to learn Hakka

- DU1: 你講故所講學老人愛學佢兜客也當難 *ngi11 gong31 gu55 so31 gong31 hog5 lo31 ngin11 oi55 hog5 en34 li55 hag2 ia24 dong24 nan11* ('So [it's] also very difficult for Southern Min natives to learn Hakka')
- DU2: 嗯嗯... *n n...* ('um um...')
- DU3: *en24 li55* 客人愛學學老較 *goi24* 啦 *en24 li55 hag2 ngin11 oi55 hog5 hog5 lo31 ka55 goi24 la* ('It is easier for us Hakka people to learn Southern Min')⁵
- DU4: 嗯嗯... *n n...* ('um um...')
- DU5: 因為電視台不時會做啊 *in24 vi55 tien55 sii55 toi11 bud2 sii11 voi55 zo55 a* ('Because there are often [Southern Min] TV programs')
- DU6: <對啊> *dui4 a* ('correct')
- DU7: <連續劇>唱歌仔就唱學老 *lian2 xu4 ju4 cong55 go24 qiu55 cong55 hog5 lo11* ('Those who sing in serials sing Southern Min')
- DU8: 係係 *he55 he55* ('yes yes')
- DU9: 故所佢兜客家人當會去講學老啊 *gu55 so31 ngai11 deu24 hag2 ga24 ngin11 dong24 voi55 hi55 gong31 hog5 lo31 a* ('So we, Hakka people, are good at speaking Southern Min')
- DU10: 學老人會講客家話就當難啊 *hog5 lo31 ngin11 voi55 gong31 hag2 ga24 fa55 qiu55 dong24 nan11 a* ('It's difficult for Southern Min people to speak Hakka!')

Subtopic 2: You are good at speaking Southern Min

- DU11: 該你學老乜當厲害喔 *ge55 ngi11 hog5 lo31 me55 dong24 li55 hoi5 o* ('You are good at Southern Min')
- DU12: 會講啦 *voi55 gong31 la* ('[I] can speak [it]')
- DU13: 會講啦係囉 *voi55 gong31 la he55 lo* ('[You] can speak [it]')
- DU14: 講講毋會當滑溜啦 *gong31 gong31 m11 voi55 dong24 vad5 liu55 la* ('[I] cannot speak it very fluently')
- DU15: 嗯嗯... *n n...* ('um um...')

⁴ Topic 1 "Speaking Hakka" was the common property shared by Subtopics 1 to 5.

⁵ Subtopic 1 "It's difficult for Southern Min people to learn Hakka" described the unfair situation that it is easier for Hakka natives to learn Southern Min but not that easy for Southern Min natives to learn Hakka. What the speakers focused on was comparing the scenarios, so it was not appropriate to classify DU3 and DU4 into another subtopic different from Subtopic 1. This part was approved by the two independent annotators in our boundary segmentation evaluation experiment.

Subtopic 3: Hakka people are poor

DU16: 故所講 **hon** gu55 so31 gong31 hon ('So')

DU17: 客家人盡衰過啊 hag2 ga24 ngin11 qin55 coi24 go55 a ('Hakka people are very poor')

Subtopic 4: Zhunan Hakka group

DU18: 毋係啦 m11 he55 la ('No, it is not true')

DU19: 一方面 hon id2 fong24 mien55 hon ('On the one hand')

DU20: 你像住躡竹南 ngi11 qiong55 ngai11 dai55 zug2 nan11 ('For example, I live in Zhunan')

DU21: 假使乜學老人 ga31 sii31 me55 hog5 lo31 ngin11 ('There are supposedly many Southern Min people')

DU22: 著 cog5 ('Yes')

Subtopic 5: Hakka people are afraid to speak Hakka

DU23: <百分之六十>个客家人 bai3 feng1 zhi1 liu4 shi2 ge55 hag2 ga24 ngin11 ('About sixty percent of the Hakka people')

DU24: 毋敢講客 m11 gam31 gong31 hag2 ('are afraid to speak Hakka')

DU25: 係啊 he55 a ('Yes, it is true')

Topic 2: Policy related to Hakka people

Subtopic 1: Minister of the Council of Hakka Affairs

DU26: 這擺斯好得<葉菊蘭>**hon** ia31 bai31 sii11 ho31 ded2 ye4 ju2 lan2 hon ('It is good to have Yeh Chu-lan this time')

DU27: 係 he55 ('Yes')

DU28: 係樣 an31 ngiong11 ('What')

DU29: <客家會主委> ke4 jia1 huei4 zhu2 wei3 ('Minister of the Council of Hakka Affairs')

DU30: 佢毋係<客委會主委> gi11 m11 he55 ke4 wei3 huei4 zhu2 wei3 ('She is not the minister of the Hakka Affairs Council')

DU31: 佢已早盡早係啦 gi11 i31 zo31 qin55 zo31 he55 la ('She was once the minister')

DU32: 係 he55 ('Yes')

Subtopic 2: Democratic Progressive Party (DPP)

DU33: 講下擺講 gong31 ha55 bai31 gong31 ('Speaking of')

DU34: 係樣民進黨愛仰仔講 **hon** an31 ngiong11 min11 jin55 dong31 oi55 ngiong31 e31 gong31 hon ('How to describe the DPP')

Understanding a conversation, for both humans and automatic systems, is to steadily obtain new and recurrent patterns of information about social interaction, linguistic content, and speaker intention. Judgments and annotations that refer to previously uttered conversational speech data are in fact indirect evidence of language planning processes inferred from the shared information between conversational partners. Nevertheless, a representational model, as we have suggested, provides a hierarchy of discourse components within a sequence of verbal interactions, and DU-initial cue words can be used to tackle the rhetorical relationships between the uttered propositions for linguistic research as well as to heuristically identify topic segmentation boundaries to enhance semantic understanding in natural language processing.

3.2 Interaction and Linguistic Patterns

Within the interaction that occurs during a conversation, new and recurrent topics and subtopics may be initiated by non-responsive or responsive actions. Tables 5 and 6 summarize the results of the DUs in terms of the three DU types, clausal, non-clausal, and fragmentary, divided into two interaction categories, non-responsive and responsive. Please note that if the discourse meaning of a DU was incomplete, that is, the speech content could not be clearly interpreted and specified, it was classified as a fragmentary meaning. The DUs whose discourse meaning could be clearly identified were divided into clausal and non-clausal meanings.

Table 5. DU types used for topic initiation

		Con. 1		Con. 2		Con. 3		Con. 4		Con. 5	
		F	M	F	M	F1	F2	F	M	F	M
Meaning	Form	Non-responsive (37)									
<i>clear</i>	<i>clausal</i>	1	2	2	4	1	7	4			3
<i>clear</i>	<i>non-clausal</i>					1	1				
<i>incomplete</i>	<i>fragmentary</i>	1	2	1		1		1		1	4
Speaker's DU proportion		33%	67%	75%	67%	100%	89%	50%		100%	100%
Meaning	Form	Responsive (18)									
<i>clear</i>	<i>clausal</i>	4					1	3	1		
<i>clear</i>	<i>non-clausal</i>			1	2				1		
<i>incomplete</i>	<i>fragmentary</i>		2					2	1		
Speaker's DU proportion		67%	33%	25%	33%		11%	50%	100%		

Table 6. DU types used for subtopic initiation

		Con. 1		Con. 2		Con. 3		Con. 4		Con. 5	
		F	M	F	M	F1	F2	F	M	F	M
Meaning	Form	Non-responsive (158)									
<i>clear</i>	<i>clausal</i>	9	11	8	9	8	17	18		7	13
<i>clear</i>	<i>non-clausal</i>	6		1	1	3	4	1		1	
<i>incomplete</i>	<i>fragmentary</i>	1	10	3	2	5	3	6	1	5	5
Speaker's DU proportion		60%	78%	48%	41%	76%	92%	48%	7%	93%	86%
Meaning	Form	Responsive (98)									
<i>clear</i>	<i>clausal</i>	10	2	7	10	3	2	23	11	1	3
<i>clear</i>	<i>non-clausal</i>		1	5	5	1		2	2		
<i>incomplete</i>	<i>fragmentary</i>		3	1	2	1		2	1		
Speaker's DU proportion		40%	22%	52%	59%	24%	8%	52%	93%	7%	14%

The overall results showed that clausal DUs were the most frequent forms in the conversational data (32/55 for topics, 170/256 for subtopics), although spontaneous conversational speech was expected to be fragmentary in terms of syntactic structure. The results support the notion that complete and coherent forms are favored in producing a locus of interaction and projecting the speaker's actions (Thompson & Couper-Kuhlen, 2005). Fragmentary DUs that had a less clear discourse meaning were actually in the minority, suggesting that when the speakers intended to initiate a new topic or subtopic, the action was to some degree already planned before the topic- or subtopic-initial DUs were produced. Our analysis also showed that the proportion of responsive DUs used for topic shifts (33%) was slightly smaller than that used for topic drifts (38%). This implied that even though we segmented the conversations into coherent topics and subtopics with different degrees of topic continuity, it was nevertheless essential for speakers to provide reactions that responded to their conversational partners' previous verbal actions.

It is noteworthy that backchannels normally refer to short utterances that are produced by the non-primary speaker or the listener when the front channel is occupied by the primary speaker, according to Yngve (1970). In the Hakka data, backchannels such as particles and short replies (e.g., 㗎 *n* and 㗎 *o*) also occurred at the topic and subtopic boundaries, and they were used to initiate a new discourse segment while responding to their conversational partner at the same time. The categorization of responsive versus non-responsive DUs contributed to an understanding of conversation interaction. For instance, the male speaker in conversation #4 started 93% of the subtopics by responding to his conversational partner's previous reaction, whereas the other speakers in the data mostly initiated subtopics without directly reacting to

their partners. The distribution of responsive DUs between the two speakers in the conversation was regarded as an indicator that represented the speaker's interaction pattern for participation behavior in a cooperative context. Currently, we are not yet in the position to claim that this is an effective indicator. Nevertheless, we have shown that in addition to linguistic patterns, the analysis of boundary DU types provided insight into the social interaction in the conversational discourses.

4. Initiation Cues in Hakka

Cue phrases in a written or spoken discourse in general refer to connectives and discourse markers that designate relevant positions for discourse segmentation and interpretation. However, to what extent discourse segments of a broader scope, such as topics and subtopics, are signaled by boundary cues with a similar function as cue phrases has not been thoroughly studied. In this study, we attempted to obtain an overview of boundary cues that were recurrently used to initiate topics and subtopics based on the annotation results of the topic and subtopic boundaries.

4.1 General Types of Boundary Cues

Cue phrases are considered pivots that deliver, change, and return linguistic messages. In particular, they are used to signal the speaker's intention for language planning and to attract the listener's attention, given that the coherence relationships between the already-expressed and the to-be-expressed propositions are intact. They may also be regarded as a type of discourse marker. Different from the conventional notion of cue phrases, strong and weak topic boundaries can be signaled by linguistic forms of different lengths, such as words, phrases, and clauses. For instance, 佢嚟你講 *le ngai11 lau24 ng11 gong31 le* ('let me tell you') and 佢會講 *ngai11 voi55 gong31* ('I would say'), produced at the topic and subtopic boundaries in the data had a function similar to that of cue phrases. We tentatively included the whole expression 佢嚟你講 and 佢會講 in the category of "empirical marker."

Table 7 lists the types of boundary cues that were used to mark topic and subtopic transitions, in which the lexical chain of a new discourse segment with either a broad (topic) or a narrow (subtopic) scope occurred. Some of the boundary cues were language-specific characteristics in Hakka and thus worthy of further investigation. In general, there was no significant difference in the distributions in terms of topics and subtopics. Noun phrases were mostly preferred for initiating topics and subtopics in Hakka, followed by connectives. The particles identified in Table 7 were not used as backchannels but instead served the discourse function of preparing the listeners for the upcoming discourse segments by signaling new information that could change the topics or subtopics. Future studies should further investigate the intonation patterns of backchannel particles and initiation cue particles to elaborate the

relationship between discourse functions and phonetic forms (Hirschberg & Litman, 1993). Empirical markers and negation markers were also identified as boundary cue types, but they did not appear as often as noun phrases, connectives, and particles.

Table 7. Occurrences of boundary cue types

Types	Topics	Subtopics
Noun phrase	16 (29%)	96 (38%)
Connective	21 (38%)	83 (32%)
Particle	13 (24%)	42 (16%)
Empirical marker	3 (5%)	26 (10%)
Negation marker	2 (4%)	9 (4%)
Total	55 (100%)	256 (100%)

4.2 Analysis of Boundary Cues

We depicted the topic development and speaker interaction in the conversations by topics, subtopics, and DUs. Utilizing this representational format, we examined initiation cues at strong and weak discourse boundaries. Different from the typical English cue phrases in Table 1, Hakka conversations do not exhibit a strong tendency to use specific groups of cue phrases that are in turn used to mark the locations of topic transitions. To gain an overview of the discourse functions of initiation cues in Hakka conversations, we conducted a pilot study. Referring to previous studies on connectives and cue phrases, we attempted to clarify the discourse functions of the boundary cues included in the results presented in Table 7. We did not implement any a priori restrictions on the length of the linguistic units, such as words or phrases, but mainly referred to recurrent patterns to specify their discourse functions. The results summarized in Table 8 are exclusively valid for our data. Herewith, we hope to provide a preliminary system of initiation boundary cues in Hakka conversations that can be further specified in more detail as well as more types of speech data.

Initiating a new discourse segment by specifying objects or qualities is a common practice in Hakka conversations. This may well explain why many noun phrases are used for topic and subtopic initiation, in addition to connectives. Most boundary cues are used for both topics and subtopics; however, in some cases that express concrete specifications of object descriptions and qualities, they do not occur at topic transition positions but are exclusively used at subtopic boundaries. We observed that boundary cues had the function of attracting the listener's attention for a topic transition. When combined with the use of lexically explicit discourse markers, that is, with a clear correspondence of function and meaning, the transition of topics and subtopics was successful and proceeded fluently within the conversations.

Table 8. Boundary cues in Hakka

Types	Functions	Boundaries	Typical Boundary Cues ⁶	
Noun phrase	Identifying time	Topic	頭擺 <i>teu11 bai31</i> ('before') 這下 <i>lia31 ha55</i> ('now')	
		Subtopic	頭擺/頭過 <i>teu11 bai31/teu55 go11</i> ('before') 這下 <i>lia31 ha55</i> ('now') 該下 <i>ge55 ha55</i> ('that time')	
	Identifying objects	Topic	佢个 <i>gi11 ge55</i> ('his')	
		Subtopic	這兜 <i>ia31 deu24</i> ('these') 該路 <i>ge55 lu55</i> ('that road')	
	Identifying places	Topic	該位 <i>gai55 vui33</i> ('that place') 你這 <i>ng11 lia31</i> ('your place')	
		Subtopic	佢个這位 <i>ga11 ge55 ia31 vi55</i> ('his place')	
	Identifying persons	Topic	你 <i>ng11</i> ('you') 佢 <i>ngai11</i> ('me') 佢 <i>gi11</i> ('he') 佢等 <i>ngai11 den31</i> ('we') 該 <i>ge55</i> ('that')	
		Subtopic	你 <i>ng11</i> ('you') 佢 <i>ngai11</i> ('me') 佢 <i>gi11</i> ('he') 佢等 <i>ngai11 den31</i> ('we') 該 <i>ge55</i> ('that') 這 <i>ia31</i> ('this') 這兜 <i>ia31 deu24</i> ('these')	
	Connective	Explication, clarification, inference	Topic	所以 <i>so31 i24</i> ('so') 因為 <i>in 24 vi55</i> ('because')
			Subtopic	所以 <i>so31 i24</i> ('so') 故所 <i>gu55 so31</i> ('so') 因為 <i>in 24 vi55</i> ('because')
		Contrast	Topic	毋過 <i>m11 go55</i> ('but')
			Subtopic	毋過 <i>m11 go55</i> ('but') 可是 <i>ko31 sii55</i> ('but')
Topic change		Subtopic	那 <i>na55</i> ('that')	
		Topic & subtopic	該 <i>ge55</i> ('that')	
Sequence		Topic	過忒 <i>go55 ted2</i> ('then')	
		Subtopic	過了 <i>go55 e31</i> ('then') 過 <i>go55</i> ('then') 然後 <i>ien11 heu55</i> ('then')	
Addition		Topic	還 <i>han11</i> ('also') 還有 <i>han11 iu24</i> ('in addition')	
		Subtopic	還 <i>han11</i> ('also')	
Concession, elaboration		Topic	其實 <i>ki11 siid5</i> ('in fact') 恁多 <i>an31 do24</i> ('so many')	
		Subtopic	其實 <i>ki11 siid5</i> ('in fact') 假使 <i>ga31 sii31</i> ('if') 恁 <i>an31</i> ('such') 敢還 <i>gam31 han11</i> ('could it still be said that...')	
Relation	Subtopic	像 <i>qiong55</i> ('like')		
Particle	Elaboration, clarification, reaffirmation	Topic & subtopic	<i>hon</i> 唉 <i>ai</i> 唉喔 <i>ai o</i> 啊 <i>a</i> 喔 <i>o</i> 嗯 <i>n</i> 噴 <i>jid</i> 諷 <i>e</i>	

⁶ The boundary cues were transcribed based on the Sixian Hakka dialect.

Empirical marker	Topic change	Topic	故所講 <i>gu55 so31 gong31</i> ('you say, so say') 僱謬你 講 <i>le ngai11 lau24 ng11 gong31 le</i> ('let me tell you') 僱會講 <i>ngai11 voi55 gong31</i> ('I would say') 就講 <i>qiu55 gong31</i> ('just speaking')
		Subtopic	僱就講 <i>ngai11 qiu55 gong31</i> ('I just say') 下擺講 <i>ha55 bai31 gong31</i> ('sometimes speaking of') 就講 <i>qiu55 gong31</i> ('just speaking') 你看 <i>ngi11 kon55</i> (‘you see’) 講 <i>gong31</i> ('saying')
Negative	Negation	Topic & subtopic	毋係 <i>m11 he55</i> ('not') 無 <i>moll</i> ('no')

5. Discussion

The semantic structure of conversational discourse needs to account for the macrostructure of the discourse and the social interaction between the conversational participants (van Dijk, 1977b). References to a given discourse referent may constantly change along the course of a conversation due to spontaneous language planning and speaker reactions. Therefore, sentence-level distinctions of topics and comments may not explicitly or effectively apply to conversational discourse descriptions (van Dijk, 1980; Asher, 2004). Our analysis of Hakka conversational data revealed that linguistic forms represented in terms of predicate-based DUs were useful in presenting the quantity and the quality of content across the subtopics. Subtopics may be more closely connected with the constructional form than a broader sense of discourse segments, such as topics defined by lexical cohesion and coherence relationships (Halliday & Hasan, 1976; Morris & Hirst, 1991; Harabagiu, 1999). Likewise, the distinction between responsive and non-responsive action types, which is important in interpreting the social interaction of participating speakers, is also more conclusive at the level of subtopics rather than topics (Hobbs, 1990; van Kuppevelt, 1995a, 1995b).

Mann and Thompson (1988) proposed rhetorical relations of propositions. If we had intended to apply the rhetorical relationship approach to decompose the content of the conversational discourses into a structured organization, we would have needed to be equipped with a sentence-comparable unit. We adopted the concept of DUs (Grosz & Sidner, 1986; Polanyi, 1995, 2005; Tao, 1996; Prévot *et al.*, 2015) to construct elementary units with which higher-level discourse segments could be built. Our results showed that DUs were effective means to link boundary cue types through discourse organization. For topic and subtopic initiation, clausal DUs are preferred (Thompson & Couper-Kuhlen, 2005). Discourse is organized based on coherence relationships that construct the “aboutness” of linguistic segments. Specifically, clauses were proven to be interactionally accessible units in our Hakka data, and our results in Tables 5 and 6 support the notion that the Hakka speakers in this study preferred clausal constructions as a linguistic strategy for topic transitions. In addition, the

discourse meaning of the DUs at the topic and subtopic boundaries also tended to be complete, suggesting that the speakers may have already completed their language planning before they produced upcoming topics.

Givón (1983) and van Kuppevelt (1995a) both proposed a hierarchical structure of discourse topics, with Givón emphasizing a horizontal relationship between the preceding discourse contexts and the current one, and Van Kuppevelt proposing that a discourse is decomposed into discourse topics, topics, and subtopics. According to Givón (1983: 12), when “lookback” is employed as a measure of topic continuity, the upper limit is 20 clauses from the previous occurrence, depending on what “the speaker makes about topic-availability to the hearer, involving the transition from ‘availability’ or ‘identifiability’ to the more neutral ‘continuity’.” It is empirically practical to rely on the principle of continuity, rather than that of discontinuity or disruption, when carrying out the task of discourse segmentation. We proposed a similar, two-level approach for describing discourse organization and topic development in Hakka conversations. The topics and subtopics were mainly identified according to topic continuity and coherence relationships. However, to achieve an understanding of the interaction within a conversation, it was necessary not only to examine the components and their relationships but also to reveal their discourse functions and the social action of the speakers. Our approach preliminarily proved useful in accounting for the linguistic characteristics of the use of DU types and boundary cues. To study speakers’ social interaction in interactive conversational speech also requires cognitive accounts that consider the intention and attention status of the conversational partners. That is, a mechanism that provides a link between cognitive states and the corresponding language production is needed (Stede, 2012; Todd, 2016). DUs, as proposed in our approach, may serve as an adequate unit for this purpose.

In the current study on Hakka conversations, we started with the discourse segmentation of the topics and subtopics by applying lexical cohesion analysis. The DUs were identified by referring to the availability of predicates, subjects, and objects according to surface structures. Following this line of data processing, we further specified the discourse functions of the boundary cues to initiate the topics and subtopics. The topic boundary cues were not limited to the specific word category of “cue phrases” but included word sequences that were recurrently used for topic and subtopic transitions. Not only were connectives and particles commonly used in spoken discourse, noun phrases that specified physical objects and qualitative properties were also preferred at the boundaries across topics and subtopics in the Hakka conversations (Tao, 2020).

6. Conclusion

Shared knowledge and semantic coherence are required for the successful execution of conversations. Dynamic changes in coherence relationships in broad and narrow senses

construct the building blocks of conversational discourse. We pointed out that predicate-based clausal accounts of DUs are an operable means of bridging information-based topic segmentation and form-based lexical processing. More studies are needed to account for linguistic properties that are directly related to social behavior, such as an effective means of making discourse segments coreferential to one another, including the use of words, sounds, prosody, and non-verbal elements. We proposed a hierarchical schema to analyze the macrostructure of conversations consisting of topics and subtopics represented in terms of DUs. Systems with more levels of discourse segments are also possible, but according to our results, the subtopics were robust units with which interactive patterns of the conversations were reflected and described. Further empirical studies examining the relationship between discourse segments, initiation cues, and phonetic forms are needed. To meet the goal of understanding and representing a conversational discourse for humans and automatic systems, it is necessary to engage in interdisciplinary collaborations to develop applicable data-driven methodologies for the automatic extraction of coherent and cohesive relationships between topics, as well as sensible mechanisms of cognitive devices that represent the intention and attention states of conversational partners.

References

- Asher, N. (1993). *Reference to abstract objects in discourse*. Kluwer Academic Press.
- Asher, N. (2004). Discourse topics. *Theoretical Linguistics*, 30(2-3), 163-201. <https://doi.org/10.1515/thli.2004.30.2-3.163>
- Asher, N., & Vieu, L. (2005). Subordinating and coordinating discourse relations. *Lingua*, 115(4), 591-610. <https://doi.org/10.1016/j.lingua.2003.09.017>
- Charolles, M. (2020). Discourse topics and digressive markers. *Journal of Pragmatics*, 161, 57-77. <https://doi.org/10.1016/j.pragma.2020.01.005>
- Cohen, R. (1984). A computational theory of the function of clue words in argument understanding. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting on Association for Computational Linguistics*, 251-258. <https://doi.org/10.3115/980491.980546>
- Das, D. (2014). *Signalling of coherence relations in discourse* (Doctoral dissertation). Simon Fraser University.
- Dias, G., & Alves, E. (2005). Discovering topic boundaries for text summarization based on word co-occurrence. In N. Nicolas et al. (Eds.), *Recent advances in natural language processing IV: Selected papers from RANLP 2005*, 187-191.
- Ferret, O., & Grau, B. (2000). A topic segmentation of texts based on semantic domains. In *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI 2000)*, 426-430.

- Fraser, B. (1996). Pragmatic markers. *Pragmatics*, 6(2), 167-190. <https://doi.org/10.1075/prag.6.2.03fra>
- Fraser, B. (2006). Towards a theory of discourse markers. In K. Fischer (Ed.), *Approaches to discourse particles* (pp.189-204). Elsevier.
- Fraser, B. (2009). Topic orientation markers. *Journal of Pragmatics*, 41(5), 892-898. <https://doi.org/10.1016/j.pragma.2008.08.006>
- Geluykens, R. (1993). Topic introduction in English conversation. *Transactions of the Philological Society*, 91(2), 181-214. <https://doi.org/10.1111/j.1467-968X.1993.tb01068.x>
- Giora, R. (1985). A text-based analysis of nonnarrative texts. *Theoretical Linguistics*, 12(2-3), 115-136. <https://doi.org/10.1515/thli.1985.12.s1.115>
- Givón, T. (1983). Topic continuity in discourse: An introduction. In T. Givón (Ed.), *Topic continuity in discourse* (pp. 1-42). John Benjamins.
- Grosz, B., & Sidner, C. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3), 175-204.
- Hakka Affairs Council (客家委員會). (2017). *Survey on national Hakka population and language basic data, 2017* (105 年度全國客家人口暨語言基礎資料調查研究), Report of Hakka Affairs Council (客家委員會研究報告).
- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.
- Harabagiu, S. M. (1999). From lexical cohesion to textual coherence: A data driven perspective. *Journal of Pattern Recognition and Artificial Intelligence*, 13(2), 247-265. <https://doi.org/10.1142/S0218001499000148>
- Hearst, M. A. (1993). *Text tiling: A quantitative approach to discourse segmentation*. Technical Report Sequoia 93/24. University of California, Berkeley.
- Hirschberg, J., & Litman, D. (1993). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3), 501-530.
- Hobbs, J. R. (1990). Topic drift. In B. Dorval (Ed.), *Conversational organization and its development* (pp. 3-22). Ablex.
- Hoey, M. (2005). *Lexical priming*. Routledge.
- Horne, M., Hansson, P., Bruce, G., Frid, J., & Filipsson, M. (2001). Cue words and the topic structure of spoken discourse: The case of Swedish men 'but'. *Journal of Pragmatics*, 33(7), 1061-1081. [https://doi.org/10.1016/S0378-2166\(00\)00044-8](https://doi.org/10.1016/S0378-2166(00)00044-8)
- Iwasaki, S. (1993). *Subjectivity in grammar and discourse: Theoretical considerations and a case study of Japanese spoken discourse*. John Benjamins.
- Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3), 243-281. <https://doi.org/10.1515/text.1.1988.8.3.243>
- Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1), 21-48.

- Nakajima, S., & Allen, J. F. (1993). *A study on prosody and discourse structure in cooperative dialogues*. 1993 Technical Report. Department of Computer Science, University of Rochester, NY.
- Polanyi, L. (1995). *The linguistic structure of discourse*. Available at: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.85.5029&rep=rep1&type=pdf>
- Polanyi, L. (2005). The linguistic structure of discourse. In D. Schiffrin et al. (Eds.), *The handbook of discourse analysis* (pp. 265-281). Blackwell Publishers.
- Prévot, L., Tseng, S.-C., Peshkov, K., & Chen, A. C.-H. (2015). Processing units in conversation: A comparative study of French and Mandarin data. *Language and Linguistics*, 16(1), 69-92. <https://doi.org/10.1177/1606822X14556605>
- Quirk, R. (1972). *A grammar of contemporary English*. Longman.
- Redeker, G. (2006). Discourse markers as attentional cues at discourse transitions. In K. Fischer (Ed.), *Approaches to discourse particles* (pp. 339-358). Elsevier.
- Schiffrin, D. (1987). *Discourse markers*. Cambridge University Press.
- Schourup, L. C. (1985). *Common discourse particles in English conversation*. Routledge.
- Stede, M. (2012). *Discourse processing*. Morgan and Claypool Publishers.
- Tao, H. (1996). *Units in Mandarin conversation: Prosody, discourse, and grammar*. John Benjamins.
- Tao, H. (2020). NP clustering in Mandarin conversational interaction. In S. A. Thompson et al. (Eds.), *The "noun phrase" across languages: An emergent unit in interaction* [Typological Studies in Language 128] (pp. 271-314). John Benjamins.
- Thompson, S. A., & Couper-Kuhlen, E. (2005). The clause as a locus of grammar and interaction. *Discourse Studies*, 7(4-5), 481-505. <https://doi.org/10.1177/1461445605054403>
- Thompson, S. A., & Tao, H. (2010). Conversation, grammar, and fixedness: Adjectives in Mandarin revisited. *Chinese Language and Discourse*, 1(1), 3-30. <https://doi.org/10.1075/cld.1.1.01tho>
- Todd, R. W. (2016). *Discourse topics*. John Benjamins Publishing Company.
- Van Dijk, T. A. (1977a). *Text and context*. London: Longman.
- Van Dijk, T. A. (1977b). Sentence topic and discourse topic. *Papers in Slavic Philology (PSP)*, 1, 49-61.
- Van Dijk, T. A. (1977c). Pragmatic connectives. *Interlanguage Studies Bulletin*, 2(2), 77-93.
- Van Dijk, T. A. (1980). *Macrostructures: An interdisciplinary study of global structures in discourse, Interaction, and Cognition*. L. Erlbaum Associates.
- Van Kuppevelt, J. (1995a). Discourse structure, topicality and questioning. *Journal of Linguistics*, 31(1), 109-147. <https://doi.org/10.1017/S002222670000058X>
- Van Kuppevelt, J. (1995b). Main structure and side structure in discourse. *Linguistics*, 33(4), 809-833. <https://doi.org/10.1515/ling.1995.33.4.809>

- Webber, B. L. (1988). *discourse deixis and discourse processing* (Technical Report MS-CIS-86-74). Linc Lab 42. Department of Computer and Information Science. University of Pennsylvania.
- Yngve, V. (1970). On getting a word in edgewise. In *Papers of the Sixth Regional Meeting of Chicago Linguistic Society*, 567-577.

應用文步分析探究言語行為一
以公共政策網路參與平臺提案文類為例

**A Move Analysis of
Communicative Acts in Petition Text on the
Public Policy Participation Network Platform**

楊惟婷*、謝承諭⁺、鍾曉芳[#]

Wei-Ting Yang, Chen-Yu Chester Hsieh, Siaw-Fong Chung

摘要

隨著科技快速發展，政府致力將資訊技術應用於創建 Join 平臺，促進人民藉由網路提案參與公共議題討論，此類文本之重要性也隨著提高。有鑑於此，本研究旨在應用文步分析，探究中文網路提案寫作架構及語言特徵，自平臺上挑選 40 篇提案文章建構語料庫，再以人工標記文章中情況、問題、解方、評價 (SPSE) 四大文步，並使用 AntConc 軟體檢索各文步中的高頻詞彙，分析其中的言語行為。研究結果整理出各文步在網路提案文章中所出現的規則，本研究結果可提供電腦自動化收集資料、分析標記文步，以及判斷訊息結構中的言語行為等功能之具體參考。

* 國立政治大學華語文碩士學位學程

Master's & Doctor's Program in Teaching Chinese as a Second Language, National Chengchi University
E-mail: 109161007@g.nccu.edu.tw

⁺ 國家教育研究院語文教育及編譯研究中心(通訊作者)

Research Center for Translation, Compilation and Language Education, National Academy for Educational Research

E-mail: chesterhugues@gmail.com

[#] 國立政治大學英國語文學系

Department of English, National Chengchi University

Email: sfchung@nccu.edu.tw

Abstract

With the rapid development of information technology, the Taiwanese government has launched the Public Policy Network Participation Platform (Join Platform), which allows citizens to start and support a petition online and voice their opinions regarding public issues. The aim of this study was to apply the method of move analysis to investigate the text structure and linguistic features of the online petition genre. In total, 40 online petition texts were collected from the website and compiled into a corpus using the AntConc application. The collected texts were then annotated with reference to the four moves of the Situation, Problem, Solution, and Evaluation textual pattern and the communicative acts in each move. The results showed that the distribution of the moves varied across the articles and that the communicative acts in each move were represented by high-frequency words. The findings of this research will thus serve as a basis for future applications, such as computerized data collection, automatic annotation of rhetorical moves, and judgment of communicative acts in texts.

關鍵詞：文步分析、問題-解方、政策論證、言語行為、Join 平臺

Keywords: Move Analysis, SPSE, Policy Argumentation, Communicative Act, Join Platform

1. 緒論 (Introduction)

文類分析(genre analysis)是國內外語言學與語言教學界熱門的研究焦點之一，過去的研究多聚焦於各領域之學術寫作(包含期刊、論文甚至演講)或是報導、社論等文類，分析語言也以英文為主，如：Hoey (2001)。英文文類分析的文獻傾向探討副詞(Charles, 2011)及抽象名詞(Flowerdew, 2008)所形成的語言特徵以及篇章架構。相較之下，中文文類分析的相關研究成果尚不及於英文，且研究的文類僅限於學術領域，尚未出現分析網路論壇文章的相關研究，因此本研究以此角度切入，試圖從應用語言學觀點建立分析架構，處理網路參與平臺提案此類型之文章。

隨著資訊科技發展，當代政府也紛紛將資訊科技應用於推動、行銷政策，建立網路平台以增進電子參與(e-participation)，這類平台遂促成民眾參與政策辯論的重要管道，同時幫助政府蒐集民意，並將民意納入政策的決策過程(陳坤毅, 2019)。不僅改變了過去民眾參與公共事務的途徑，也改變了政府推行政策、解決問題的方式。本研究探討之公共政策網路參與平臺(以下簡稱 Join 平臺)，期望提供民眾公平參與政策的機會，落實網路公民參與。Join 平臺提供「提點子」機制，由民眾針對公共政策提供不同見解或是建言，民眾提案須先經過行政機關檢核通過後才能進入附議程序，並且在 60 天內達到超過 5000 人附議，即可成案，而對應提案的政府相關部門需作出正式回應。民眾依據個人需求及關注議題在 Join 平臺提案，尋求獲得廣大民眾附議支持，最終期待獲得政府正式回應與實際解決。民眾往往將自身訴求寫入提案內容，或是對特定議題、群體提出特定請

求。因此在寫作手法上，提案內容必須明確點出現況所出現的問題，同時提出具體的解決方案，以證明提案的必須性和迫切性。

儘管這樣的新興文類在現代社會治理當中，扮演了越來越重要的角色，但是在學界，尤其是語言學界，卻較缺乏相關的分析。其中較為值得注意的是 Hagen 等人(2016)的研究。該團隊運用計算機語言學方法，針對英文電子參與平臺上文章進行分析，調查了提案型文類中潛在的語言模式，並探討語言特徵和語意特徵是否能有效吸引民眾參與電子討論。研究結果發現，使用極端語言的提案並不會吸引民眾討論，反而是與重要社會事件相關的提案容易引發廣泛討論。作者們亦指出，重複某些詞彙或是眾所熟悉的主題有助於增加提案的急迫性。然而，Hagen 等人(2016)的研究以及相關之文獻，缺乏從文類架構的角度來分析何為提案類型文章最有效、具說服力的寫作技巧，因此，本研究試圖以「問題—解方」的架構(Hoey, 2001)來分析此類提案文章。

以「問題—解方」(problem-solution pattern)為架構的寫作手法經常應用於學術寫作以及新聞寫作上，使用此架構的特色在於作者能在「問題」文步中，善用語言及修辭知識凸顯將探討之議題的價值、可討論性及解決方案提出的可能性；在「解方」文步中提出全文的核心重點(Flowerdew, 2003)。過去研究探討的方向多為語言、語意特徵，而本研究採用 Hoey (2001)所提出章節架構(micro-structure)中「問題—解方」的理論為研究框架，以「文步」(moves)為單位，分析平臺文章的訊息架構，並且深入探究文步中常見的語言特徵及語言模式(patterns)，以及細緻討論語言模式所產生的各種言語行為(speech/communicative act) (Ali, 2013)，由於文步內本來就嵌入以交際為目的的言語行為(Hyland, 1990)，因此探究文步中的言語行為能幫助研究者更適當地判斷文步。並有助於應用研究成果於自然語言處理等應用的層面。

綜上所述，本文提出以下三點研究問題：

- 一、Join 平臺上提案型文類寫作是否應用「問題-解方」架構？
- 二、Join 平臺上提案型文類寫作的文步分布及特性為何？
- 三、提案型文類的文步中所含的語言特徵、模式及言語行為為何？

2. 文獻回顧 (Literature Review)

2.1 「問題-解方」文步分析 (The Problem-Solution Pattern)

在文類分析(genre analysis)的研究中，文類(genre)被視作是一個具有特定溝通功能與任務的文本類型(Bhatia, 1993; Swales, 1990)。而要達到該文類的功能與任務，文本通常都有特定的修辭成份與結構，這些成份與結構通常具有一定的序列性，也有各自的修辭功能與語言特徵，學者稱之為「文步」(move)。分析文類特屬的文步，不僅能讓我們更加了解該文類的特性與本質，研究成果更能應用在語言教學與自然語言處理等等應用的層面(Swales & Feak, 2012; Heffernan & Teufel, 2018; Ratanakul, 2018; 黃冠誠等人, 2014)。

在文獻當中，文步分析研究最為廣泛且深入，即學術寫作相關的文類。例如在學術英文研究中，最有名的一個文步結構，便是由 Swales (1990)所提出的「創造研究空間」

(Create A Research Space, CARS)模型。Swales (1990)在研究了不同領域學術論文的緒論章節後指出，儘管領域間稍有差異，但大多數學術論文的緒論部份，都會具有三個文步：(1)界定研究範圍(establishing a territory)、(2)建立研究利基(establishing a niche)、(3)佔據研究利基(occupying a niche)。這些文步各自具有常用的語言表達，也可以再細分為更小的子文步(step) (Swales, 1990)，而這個文步結構在不同領域論文的緒論部份，也因為領域研究與修辭的傳統，而有些許的變化與差異(Samraj, 2002)。

除了「創造研究空間」這個模型外，另一個最常被研究與分析的文步結構，便是本文主要探討的「問題-解方」架構(Problem-Solution pattern) (Hoey, 1983; 2001)。根據英國語言學家 Michael Hoey (1983, 2001)的說法，「問題-解方」架構通常具有四個元素：(1)情境(situation)、(2)問題(problem)、(3)解方(solution)或回應(response)，以及(4)評價(evaluation)，因此又常被簡稱作 SPSE 或 SPRE 架構。「情境」這個文步呈現與主題相關的人事物以及背景資訊；「問題」描述需要處理的狀況，包含阻礙與難題，或是現況的缺點；「解方」或「回應」的部分，則是提出解決或處理問題的方法；最後「評價」指的是針對解方或是回應的優缺點提出的分析與評論，如果多於一個解方，則通常會比較其優劣。

相似於先前所提到的「創造研究空間」模型，「問題-解方」架構具有特定的常見成份，各個成份具有其獨特的語言特徵。例如 Hoey (2001)提到，「問題-解方」這個修辭結構的語言標記，常常出現在詞彙的層次，最常見是詞義直接與這個文步結構相關的名詞，如 *problem* 和 *solution*。其次像是有評價功能的字詞或片語，無論是本身就直接表達評價的，如 *unfortunately*，或是語義能間接透露意見的，如 *have no money*，也都可能在寫作中用來達到這個功能。Flowerdew (2008)針對學術報告的研究，則分析了「問題-解方」架構常見的詞彙，包含抽象名詞，如 *problem(s)*、*solution(s)*、*implementation* 與 *recommendations*，以及特定動詞的完成式或過去分詞型，如 *recommended* 和 *proposed*。另一方面，Charles (2011)指出，英文中的結果副詞(adverbials of result)，如 *thus*，和轉折副詞(adverbials of contrast)，如 *however*，也常在學術論文中一起使用來標記「問題-解方」架構中的不同文步。有別於上述研究，Ali (2013)探討了商業新聞文本「問題-解方」架構中的言語行為(speech/communicative act)，她發現，問題和解方文步之間，存有顯著的差異，而在各文步中，也都有更細微的言語或溝通行為，儘管並非是一對一的關係，例如問題文步中常有告知(informing)、評價(evaluating)、預測(predicting)、解釋(explicating)和退讓(making concessions)等行為，而解方文步裡則常有直述句(direct statement)、祈使句(imperative)、與評價結合的祈使句等等。

然而，不同於「創造研究空間」模式多使用在寫作學術文章的緒論上，「問題-解方」架構可以應用在許多不同文體的寫作與分析當中(Hoey, 2001)，包含學術報告(Flowerdew, 2003; 2008)、科學研究論文(Charles, 2011; Heffernan & Teufel, 2018)、新聞報導寫作(Ali, 2013; Belmonte, 2009; Ratanakul, 2017)，甚至是廣告(Hoey, 2001)與學術演講(Ratanakul, 2017)，從以上條列之文獻足見「問題-解方」這個文步架構的重要性以及應用範圍之廣泛。「問題-解方」架構的研究更指出，即使是在同一個文類中，許多因素也會影響這個文步

結構的呈現與序列，例如作者的專業程度(Flowerdew, 2003)、學科領域(Charles, 2011; Heffernan & Teufel, 2018)、語言變異與地區文化(Ali, 2013; Ratanakul, 2018)等等，由此可知，「問題-解方」這個文步架構仍有許多值得探索與發掘的主題。

2.2 文步中的言語行為 (Communicative Acts in Moves)

言語行為理論(Speech Act Theory)是由英國哲學家 John Austin 所提出，所謂的言語行為，指的是利用語言以實現說話者或作者溝通意圖與目的的行為，該理論同時強調語境對於理解言語行為的重要性。Austin (1962)將言語行為的意義由表層至深層分成：言內行為(locutionary act)、言外行為(illocutionary act)和言後行為(perlocutionary act)三層次。言內行為指說話者所使用的語言、文字；言外行為指說話者藉由語言表達內心真正的想法；言後行為指聽話者接受、理解說話者的言外行為後所做出的反應。完成一個有效的言語行為需要仰賴說話者與聽話者雙方對這三層面的理解。

Searle (1976)則依據說話者的意圖與說話內容分成直接言語行為(direct speech act)和間接言語行為(indirect speech act)。直接言語行為使用行事動詞(performative verbs)直接表達請求，道歉，答應，宣布，警告，命令，提議等意圖；間接言語行為則無法直接從語言表面看出真正意圖。例如：「請打開窗戶。」句子中使用行事動詞直接表達請求意圖，同時，在相同語境下，說話者也可以說「教室裡有點熱。」間接傳達請求意圖。

言語行為適用於分析溝通行為與文類，從微觀的角度探究文步中細微的言語行為。由於文步內本來就嵌入以交際為目的的言語行為(Hyland, 1990)，因此探究文步中的言語行為能幫助研究者更適當地判斷文步。Ali (2013)探討了商業新聞文本「問題-解方」架構中的直接言語行為，稱之為溝通行為(communicative act)，並發現不同的文步，各自具有不同的直接言語/溝通行為特徵：如「問題」文步中常有告知(informing)、評價(evaluating)、預測(predicting)、解釋(explicating)和退讓(making concessions)；「解方」文步裡常有直述句(direct statement)、祈使句(imperative)、與評價結合的祈使句等等的言語/溝通行為。

由此可知，儘管文步和言語行為之間沒有一對一的對應關係，但是只要能掌握某一文類裡各文步的特色、言語行為等，將有助於提升文步判斷的準確度，甚至可進一步用來預測相同類型文章的文步。

綜合上述所提的文獻，本研究欲探討 Join 平臺上提案型文類寫作如何應用「問題-解方」架構，解決、處理及評價議題，探討文中的文步分布與特性，更進一步分析各文步中的語言特徵、語言模式及文步與言語行為之間的對應關係。

3. 研究方法 (Methodology)

3.1 資料來源 (Data Source)

本研究欲收集提案類型文章，分析其中所包含的文步以及言語行為，因此以「公共政策網路參與平臺」(Join 平臺)為研究個案，收集民眾成功提案且受政府回應之內容。其中「提點子」這項功能即可讓公民根據己身訴求，自行擬定提案，經過民眾附議、電子連

署等檢核階段，即可成案，且依規定，政府必須在一定的時間內回應已成案的內容。

提案的主題多元，涵蓋醫療、教育、社會、法律等等議題，且依照提案格式之規定，書寫內容必須包含兩個「提議內容或建議事項」及「利益與影響」兩大部分。下圖為 Join 平臺介面「提點子」的網頁截圖。



圖 1. Join 平臺介面
[Figure 1. A Snapshot of the Join Platform]

3.2 資料收集 (Data Collection)

本研究利用 Join 平臺上的公開資訊，整理自 2017 年 10 月至 2021 年 6 月間，政府已經正式回應之成案資料共 148 則，再進行字數篩選。由於本研究期望探討成案文章中的寫作架構與語言特徵，並且使用文步分析探究段落寫作特色，因而設定書寫內容「提議內容或建議事項」及「利益與影響」的字數各為 250 字，剔除字數不足、無法成段的文章，也剔除文章中所夾帶的網址連結與圖片連結，最後共收集 40 則提案內容建構語料庫。

提議內容或建議事項

近年來，在基層教練努力與家長們的支持下，臺灣基層足球發展愈趨興盛，各項賽事參賽隊數也屢創新高。然而，基層足球人口上升，所伴隨來的最直接困境即為各縣市現有場地規畫之不足。

有於足球發展之弱勢以及公部門對於足球場地設施往往受十一人制球場的想像限制，無論是都會區運動設施興建過程中對於足球場地規劃的排斥，致使公用足球場設施缺乏。此外，近年政府在體育設施上又多集中於學校，又衍生各校因管理需求難以租借使用狀況。然而，運動場地之設置不應該只是零合遊戲，而足球場地亦不僅只有十一人制之選項、五人、八人以及非正式之複合型兩用球場等社區型球場在日本及歐洲亦為相當常見之社區球場形式。

參考德國在 1998 年世界杯和 2000 年歐洲錦標賽失利後，德國對於小型足球場興建之政策，以及 2006 年世界盃後，德國足協與官方共同制定之十年計劃「FUSSBALL IST ZUKUNFT」中，為提供青少年參與足球運動之可能，興建千座小型社區球場。此計劃不只提高了德國民眾足球活動的機會與熱情，更加強學校與俱樂部之間的合作可能，並降低了女子足球與移民背景的兒童融入社會的障礙。同時，社區球場的設置，也成為後繼德國於 2014 年世界盃奪冠、2016 年歐洲杯四強基層育成的基石。

綜上，借鑑前述德國經驗與台灣目前足球推廣實務困境，為提升國內足球風氣與參與可能，爰提出本次台灣版之「千足計劃」倡議，期待透過中央政府之鼓勵與試點，打破過去多在學校設置足球場之校園本位、翻轉足球場只能大型或專用的刻板印象，以千座場地為目標，透過設置社區型足球場或翻修既有場地成為複合式使用球場(籃、足球兩用，或檯、足球兩用等)並規劃試辦計劃，增加我國運動空間使用之可能。

圖 2. 收集至建構語料庫的提案範例
[Figure 2. An Example Text in the Petition Corpus]

圖 2 為符合字數篩選的範例文章截圖以及圖 3 為不符合字數篩選的範例文章截圖。

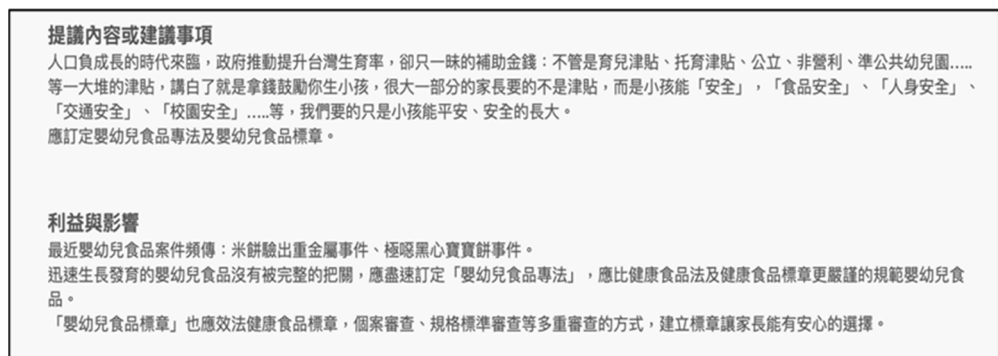


圖 3. 因字數不足而未選入語料庫的範例

[Figure 3. An Example of Texts Not Included Due to Word Counts]

3.3 資料處理 (Data Analysis)

篩選後的文章以 Excel 檔案儲存。首先，第一步使用 CKIP CoreNLP 將文章斷詞。第二步進行文步之判斷與分類，以人工標記「情況」、「問題」、「解方」、「評價」四大文步，文步判斷標準參照 Hoey (2001:123) 之分類標準和舉例：(一)、「情況」文步：通常位於文章開頭作為引言，簡介問題起源、給予框架。例子：我曾是一位英語老師(I was once a teacher of English Language.)。(二)、「問題」文步：通常接續著情況文步，明確指出目前遇到的困難或是急需解決的難題。例子：有一天學生們來找我因為他們不會寫名字(One day some students came to me unable to write their names.)。(三)、「解方」文步：必須清楚說明將使用什麼方法解決問題。例子：於是我教導他們分析文本(I taught them text analysis.)。(四)、「評價」文步：總結從中獲得效益，若有不足之處也須提出，作為未來展望。例子：現在他們都能寫小說(Now they all write novels.)。

表 1 以「名正言稅 請支持記帳士正名為稅務士！」一文作為文步分析釋例。統計各文步在文章中的佔比。

表1. 文步分析舉例
[Table 1. Examples of Each Move]

文步	舉例
<p>情況</p> <p>位於文章開頭作為引言，簡介問題起源</p>	<p>台灣面向國際，從正名開始舉觀各國，美國、德國、韓國、日本等先進國家皆已於1980年代更名為「稅務士」或「稅理士」，中國大陸更直接稱為「註冊稅務師」。</p>
<p>問題</p> <p>明確指出目前遇到的困難或是急需解決的難題</p>	<p>記帳士遲遲未獲得正名，已無形中窄化台灣的國際空間發展。...全國有近9000名稅務從業人員，每日兢兢業業。可說是台灣的每一個中小企業與商家的背後，都有記帳士的身影...然2009年經大法官釋字第655號解釋，記帳士執業內容顯然已超過「記帳」二字的涵蓋範圍，且記帳士專業資格乃經國家考試院賦予專業認證，「記帳士」的職業名稱顯然已與其執業範圍明顯不符。職業名稱長期誤導社會認知...身為記帳士主管機關的財政部，卻遲未針對大法官之解釋提出相應的解決方案。</p>
<p>解方</p> <p>清楚說明將使用什麼方法解決問題</p>	<p>我們提案：1・要求財政部研議「記帳士法」修法案，將「記帳士」正名為「稅務士」，將「記帳士法」正名為「稅務士法」，並提交行政院院會。2・要求行政院院會慎重審理記帳士法修法提案，並將該法案列入行政院優先法案。</p>
<p>評價</p> <p>總結從中獲得效益</p>	<p>記帳士長期默默耕耘，為在地產業的支持力量、新創事業最佳夥伴企業委託記帳士處理稅務問題，需要仰賴其專業度以及業務。...倘若不仰賴記帳士，根本無法滿足台灣中小企業與商店家的需求。...正名為稅務士後將有利於建立正確的職業觀念，吸引學生報考相關系所並取得資格考試投入職場，健全國家稅收稽徵的運作。</p>

完成文步分類後，進行言語行為之標記。本研究應用 Ali (2013)所提出之直接言語行為標籤，分別為：陳述(informative)、評價(evaluative)、凸顯(assertive)、解釋(explicative)、預測(predictive)、祈使(imperative)、比較(comparative)、對比(contrastive)、讓步((concessive)、分類(classification)、提問(questions)，分析文步中的言語行為。下表為 Ali (2013)及本研究所使用之直接言語行為標籤及其釋義。

表 2. Ali (2013) 直接言語行為標籤

[Table 2. Definitions and Examples of the Communicative Acts Based on Ali (2013)]

直接言語行為	定義與例句
陳述	客觀描述某事件、行為或思想。 例句：但第一次酒駕犯罪者，如造成人員受傷或死亡，就必須加鞭刑 1 鞭，開始起算。
評價	給予判斷、評論、理由或是意見。 例句：根本是喪權辱國的不平等條約。
凸顯	刻意顯現特定立場或假設。 例句：精神病患所需之醫療協助，不會因為時空改變而改變。
解釋	使用範例、公式說明概念。 例句：例如離岸流或陡降型海岸的潛在危險之介紹。
預測	提出理論上的假設以及可能性。 例句：臺灣能贏得更多國際的掌聲及尊重。
祈使	提出建議、禁止、建議或是要求。 例句：請求行政院相關部會於內部訂立相關事件程序規範，改善現況。
比較	將不同的事實及情況做比較，點出相同與相異之處。 例句：與潛在暴力風險相比，到底是否符合最大利益應為大眾所討論。
對比	將兩個反面的事實及情況做對比。 例句：缺乏與時俱進的法規護體，小蝦米難抗大鯨魚。
讓步	將對立的事實置於假設的因果關係，其中一項事實可能出現不足的先決條件或是意料之外的結果。 例句：除了暴力本身應有的懲罰，在醫療場所的暴力行為造成的影響絕對比路邊械鬥來得嚴重許多。
分類	依照關係將內容做出排序及次序分類。 例句：項目有風浪板、水上摩托車、獨木舟、風箏衝浪等。
提問	提出疑問、懸問或是反問。 例句：難道國家希望再一次看見悲劇發生嗎？

本研究應用 AntConc 語料庫軟體檢索文步中的所含有的高頻詞 (<https://www.laurenceanthony.net/software/antconc/>)，分析提案文章中所應用的言語行為。圖 4 為使用 AntConc 軟體檢索詞彙介面截圖。

1	作用(Na) · (COMMACATEGORY) 我們(Nh) 呼籲(VE) 政府(Na) 藍(D) 參考(VC) 歐洲(Nc) 先進(VH) 國家(Na)	解決方法 soluti
2	潛在(A) 危險(VH) 不了解(VK) · (COMMACATEGORY) 惹(Cbb) 藍(D) 在(P) 海岸(Nc) 建置(VC) 警告(Na)	解決方法 soluti
3	本位(Na) · (PAUSECATEGORY) 翻轉(VAC) 足球場(Nc) 只(Da) 藍(D) 大型(Na) 或(Caa) 專用(VI) 的(DE)	解決方法 soluti
4) 處(Nf) 幼兒(Na) · (COMMACATEGORY) 如此(VH) 才(Da) 藍(D) 實踐(VC) 與(Caa) 達成(VC) 行政院(Nc)	解決方法 soluti
5) 服務(VC) · (COMMACATEGORY) 相信(VK) 親主(Na) 都(D) 藍(D) 接受(VC)) (PARENTHESISCATEGORY) 搭乘(VC) 書(D) 攜帶(解決方法 soluti
6) (PARENTHESISCATEGORY) 實踐(Na) · (COMMACATEGORY) 藍(D) 提高(VC) 存活率(Na) [(PARENTHESISCATEGORY) 1(Neu)]	解決方法 soluti
7) · (PERIODCATEGORY)3(Cbb) 相關(VH) 原則(Na) 期望(VK) 藍(D) 有(V_2) 嚇阻(Nv) 作用(Na) 對於(P)	解決方法 soluti
8	na)) (PARENTHESISCATEGORY) · (COMMACATEGORY) 她(D) 藍(D) 有利於(VK) 國家(Na) 永續(VH) 發展(VC) · (解決方法 soluti
9	(Nc) 針對(P) 威脅(Na) 神經(Na) 醫學(Na) 藍(D) 有效(VH) 葬地(VA) 整體(Na) 醫療(VC)	解決方法 soluti
10	(Na) 著(Nc) 提到(VE) vegan(FW) 純素(Na) 藍(D) 減低(VC) 大部分(Nega) 的(DE) 優化(VHC)	解決方法 soluti
11	CATEGORY) MRI(FW) 檢查(VE) · (COMMACATEGORY) 才(Da) 藍(D) 減少(VHC) 勝過(Na) 及(Caa) 其他(Nega)	解決方法 soluti
12	(Na) 共同(A) 推動(VC) 的(DE) 法律(Na) 藍(D) 發揮(VJ) 作用(Na) · (COMMACATEGORY) 我們(Nh) 呼籲(解決方法 soluti
13) 的(DE) 共識(Na) · (COMMACATEGORY) 使(VL) 訓練(Na) 藍(D) 真正(D) 滿足(VHC) 新進(A) 中醫師(Na)	解決方法 soluti
14	(A) 中醫師(Na) 的(DE) 臨床(A) 能力(Na) 藍(D) 真正(D) 得到(VJ) 提升(VC) 以(P)	解決方法 soluti
15) · (COMMACATEGORY) 五(Neu) 年(Nf) 內(Ng) 未(D) 藍(D) 確定(VK) 應(VC) 納稅額(Na) 者(Na) · (解決方法 soluti
16	(Nega) 非(D) 一(Neu) 篇(Neu) 日(Nf) 藍(D) 解決(VC) · (PERIODCATEGORY) 然而(Cbb) · (COMMACATEGO	解決方法 soluti
17) 處罰(Na) · (COMMACATEGORY) 此(Nep) 舉(Na) 也(D) 藍(D) 讓(VL) 托嬰(VB) 或(Caa) 裕母(Na)	解決方法 soluti
18) 的(DE) 施政(VA) 理念(Na) · (COMMACATEGORY) 亦(D) 藍(D) 讓(VL) 家長(Na) 感受到(VK) 中央(Nc)	解決方法 soluti
19	(Na) 的(DE) 車安(Na) 中心(Nc) 如何(D) 藍(D) 負責(VL) 國人(Na) 的(DE) 車輛(Na)	解決方法 soluti
20	之(DE) 探討(VE) · (COMMACATEGORY) 以順(VK) 將來(Nd) 藍(D) 達到(VJ) 整齊(VJ) 全民(Na) 健康(VH)	解決方法 soluti
21	(Na) 種(Na) 改用(VC) 米糠(Na) 也(D) 藍(D) 避免(VE) 造成(VK) 民眾(Na) 情緒(Na)	解決方法 soluti
22	(DE) 營養(Na) 與(Caa) 體力(Na) 才(Da) 藍(D) 配合(VC) 疾病(Na) 的(DE) 各(Nes)	解決方法 soluti

圖 4. 使用 AntConc 軟體檢索高詞頻詞彙

[Figure 4. Concordance Lines of High-frequency Words Generated by AntConc]

下圖 5 為本研究之架構流程圖。

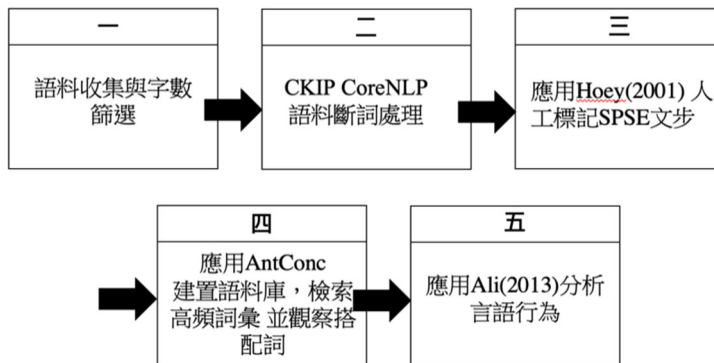


圖 5. 研究架構流程圖

[Figure 5. Flow-chart of the Research Process]

4. 研究結果與分析 (Results and Analysis)

本研究收集 40 篇提案文章，經 CKIP 斷詞後，總詞數共 26,515 詞，並使用 Excel 進行標記與統計，所得數據見下表 3。如表中所示，四大文步所含的詞數分別為：「情況」計 1,806 詞；「問題」計 7,347 詞；「解方」計 4,848 詞；「評價」計 12,514 詞，共計 54,388 詞。其中詞數最多的文步為「評價」，佔總字數 47.20%；次多的文步是「問題」，佔總字數 27.71%；再者是「解方」，佔總詞數 18.28%；而「情況」文步的詞數是四大文步中最少的，只佔總詞數 6.81%。

至於篇數方面，透過統計後發現 40 篇提案中皆存在的文步為「評價」，其餘文步則不會出現在每一篇提案中。「問題」文步出現了 33 篇；「解方」文步出現了 37 篇；而「情況」文步只出現在 30 篇提案文章裡。

表3. 提案文章中的文步比例
[Table 3. Distribution of SPSE Moves in the Corpus]

文步	提案文章(共 40 篇)		
	詞數	百分比%	文步出現篇數
情況	1,806	6.81%	30
問題	7,347	27.71%	33
解方	4,848	18.28%	37
評價	12,514	47.20%	40
共計	26,515	100%	

本研究應用 Ali (2013) 直接言語行為分類進行標記，所得出的數據如表 4。表中所示，在各個文步中，陳述和凸顯是此文類中經常使用的言語行為，用以客觀說明某件事實；而「問題」文步最常使用凸顯行為，推測意圖為加強呈現問題的嚴重性；「解方」文步中大量出現祈使句，提出建議或是禁止；「評價」文步會出現判斷、評論、理由或是意見。而在 Ali (2013) 的研究中卻發現「問題」傾向使用陳述、預測及讓步這幾種言語行為以凸顯問題的急迫性，推測可能是因為 Ali 所使用的分析資料為新聞文本造成的差異。

表4. 各文步中直接言語行為
[Table 4. Communicative Acts in Each Move]

直接言語行為	情況		問題		解方		評價	
	次數	%	次數	%	次數	%	次數	%
陳述	26	27.7%	55	19.6%	24	11.9%	89	16.9%
評價	10	10.6%	50	17.9%	24	11.9%	97	18.4%
凸顯	22	23.4%	71	25.4%	22	10.9%	67	12.7%
解釋	16	17.0%	28	10.0%	22	10.9%	52	9.9%
預測	3	3.2%	10	3.6%	12	5.9%	56	10.6%
祈使	8	8.5%	21	7.5%	75	37.1%	86	16.3%
比較	5	5.3%	20	7.1%	11	5.4%	26	4.9%
對比	1	1.1%	6	2.1%	1	0.5%	19	3.6%
讓步	0	0.0%	7	2.5%	8	4.0%	13	2.5%
分類	1	1.1%	2	0.7%	2	1.0%	9	1.7%
提問	2	2.1%	10	3.6%	1	0.5%	12	2.3%
共計	94	100%	280	100%	202	100%	526	100%

除了檢視言語行為，本研究也透過 AntConc 得出各項文步中頻率(frequency)最高的前 30 詞，數據如下表 5 所示。詞表標示各文步中詞頻最高的前 30 詞以及詞頻，由於各文步所涵蓋詞數並不相同，若是使用詞頻分析可能會導致分析結果有所偏差，因此本研究更傾向從詞的排序(ranking)尋找各個文步的高頻詞，分析詞彙所蘊含的言語行為。雖然 Ali (2013)提到，每個書寫者所運用的書寫習慣、用字遣詞並不一致，導致歸納詞彙特徵(lexical features)時有一定的困難度，但是高頻詞搜尋仍有助於成為本研究判斷言語行為的依據。

表 5. 各文步中頻率最高的前 30 個詞
[Table 5. Top 30 Frequent Words in Each Move]

情況		問題		解方		評價		全文	
詞彙	詞頻	詞彙	詞頻	詞彙	詞頻	詞彙	詞頻	詞彙	詞頻
的	50	的	266	的	138	的	541	的	995
之	23	之	89	之	88	之	140	之	340
及	18	是	58	及	51	在	109	及	225
是	18	與	58	或	35	及	101	在	197
在	17	年	56	應	34	有	95	與	185
專利	16	及	55	政府	32	是	84	有	177
與	16	在	53	並	31	不	83	是	170
年	14	不	48	等	31	與	82	不	155
或	12	有	48	與	29	為	62	或	137
來	10	或	39	以	27	對	60	年	131
連結	10	一	36	志工	27	讓	56	為	118
以	9	等	33	有	26	醫療	56	者	115
已	9	者	32	婚姻	25	也	53	等	112
為	9	人	30	於	24	者	53	以	110
其	8	而	30	者	23	而	53	對	109
有	8	也	29	能	23	或	51	政府	102
藥品	8	都	29	人員	22	以	50	而	102
障礙	8	台灣	27	為	22	政府	49	並	101
動物	7	對	27	當事人	21	能	49	也	101

國家	7	動物	26	學校	20	動物	44	能	99
機關	7	能	26	年	20	並	43	醫療	97
等	7	中	25	不	19	人員	43	一	92
者	7	為	25	在	18	台灣	43	應	90
一	6	以	24	後	18	年	41	人員	89
並	6	人員	23	對	17	等	41	台灣	87
各	6	會	23	由	17	一	40	讓	86
文化	6	更	23	第	17	可	40	其	83
機構	6	並	21	訓練	17	更	39	動物	82
相關	6	民眾	21	醫療	17	人	38	於	75
而	6	飲食	21	也	16	其	38	更	75

4.1 情況文步 (Situation)

Hoey (2001)指出在不同類型的語境下，有些文步是必須存在，而有些是可有可無的，如「情況」文步。Flowerdew (2003)重探「問題-解方」理論架構，提到「情況」文步通常位於文章開頭作為引言，簡介問題起源。從詞表中發現此文步中最常出現的動詞是具有判斷功能的「是」。「是」字句是十分常見且特殊的中文句型，在王錦慧和何淑貞 (2010)所著一書中將「是」字句歸納為華語中特殊句型，「是」作為一個判斷詞表示肯定，並不表示動作，且當「是」作為句子的中心動詞時，賓語會引導或是歸納出某種情境。如例(1)與(2)：

- (1) 『山豬吊』是一種以金屬材質繩索並以續壓式彈簧裝置束綁的陷阱，與捕獸鉗不同...
- (2) 近年來，發生攻擊監所管理員及法警事件導致受傷流血，比例更是逐年大量增加，從事都是危及生命的工作。

提案者也經常使用「在...年」、「近年來」、「逐年」、「歷年」、過去某個特定時間點或是過去曾發生過的案件起頭，描述問題起因與來龍去脈，因此「+年」詞組在此文步中出現的次數非常多，如例(3)與(4)：

- (3) 農曆七月義民祭的神豬祭祀常被誤以為是客家文化，但早在2011年，就有上百位客家人率先發起連署...

- (4) **近年**來，在基層教練努力與家長們的支持下，臺灣基層足球發展愈趨興盛，各項賽事參賽隊數也屢創新高。

除了點出過去某個特定的時間以外，提案者也經常使用「已」作為比較用途，由於Join 平臺文章的書寫目的以「管制政策」為多數，期望藉由提案修改現行不合時宜的法律、政策，藉由在「情況」文步中比較他國類似的案例，有助於加強凸顯提案的必要性與緊急性；同時，提案文章往往大量引用法律條文，而法律條文使用大量書面語書寫，如「者」、「而」、「為」...等，其中「者」常用來表示人或事物的代稱，經常出現在法律、報章雜誌、社會科學領域等相關文章裡，代指特定對象，「+者」經常出現於法律條文中。

- (5) 另外，經常實施醫療暴力**者**為病患的家屬，也應考慮列入註記的選項。
- (6) 第 221 條對於男女以強暴、脅迫、恐嚇、催眠術或其他違反其意願之方法而為性交**者**，處三年以上十年以下有期徒刑。

即使每個書寫者的寫作風格、用字遣詞不一致，導致在歸納詞彙特徵時產生困難度，但是仍不可抹滅歸納高頻詞在判斷言語行為的重要性，這些特定的語言表達方式成為讀者瞭解作者寫作意圖的信號字(Vestergaard, 2000)。圖 6 為本研究歸納「情況」文步中，表達各項功能的信號字，並且應用 Ali (2013)提出的 11 項言語行為標籤做為分類。



圖 6. 「情況」文步高頻詞與對應之言語行為
[Figure 6. High-frequency Words and the Corresponding Communicative Acts in the Situation Move]

然而，並非每一篇提案都會出現「情況」文步，本研究發現「情況」文步在 40 篇文章中出現 30 篇，顯示在此類型的語境下，「情況」屬於可有可無的文步。Hoey (2001)也指出依照不同的語境，有些文步是必須存在，而有些是可有可無的，如「情況」文步。在語料中顯示若是提案者選擇開門見山的方式點出急需解決的問題點，則不會使用「情況」文步開頭，轉而選擇使用「問題」文步起頭，如例(7)：

- (7) 針對準公共化政策居家托育人員部份，最大的爭議在於『約定契約』及政府用『支付 6000 元』購買托育服務，衍生托育契約與保親關係問題…。

4.2 問題文步 (Problem)

「問題」文步通常接續著情況文步，接續說明在哪些情況下，產生了困難或是急需解決的難題，因此一開始也會使用「在」指出前面所提到的特定情況，(Flowerdew, 2003)指出提案者傾向使用因果關係的修辭技巧寫作，描寫問題產生的前因後果。但是在本研究的語料中「因為」、「所以」等標示因果關係的詞卻並未成為高頻詞，同樣地，「問題」、「難題」等相關抽象名詞也未成為本文步的高頻詞。反而是其他帶對比意味的高頻詞，如：「卻」、「而」、「但」、「未」、「不」大量出現，用以凸顯現況與理想狀況之對比，以及尚需改進之處，而依據 Charles (2011)這些帶有對比意味的高頻詞可成為判斷「問題」文步的重要信號字。如例(8)。

除此之外，「台灣」、「政府」這兩個詞除了出現在「問題」文步中，也出現在其他文步，可能代表此文類的特色，因為提案類型的文章的寫作目的是期待與政府互動，最終得到政府回應。但是「台灣」、「政府」在「問題」文步裡的排名比較前面，是較為獨特的現象，且提案人傾向使用反問或是直接否定的句式，表達對台灣政府的不滿，如例(9)。最後，此文步所含之常見言語行為及信號字歸納在圖 7。

- (8) 各國在美方壓力下引入專利連結制度時，至少都換得與美國簽訂自由貿易協定(FTA)或加入區域經貿協定，台灣卻不然。
- (9) 但實際上政府有看到這些歧視嗎？

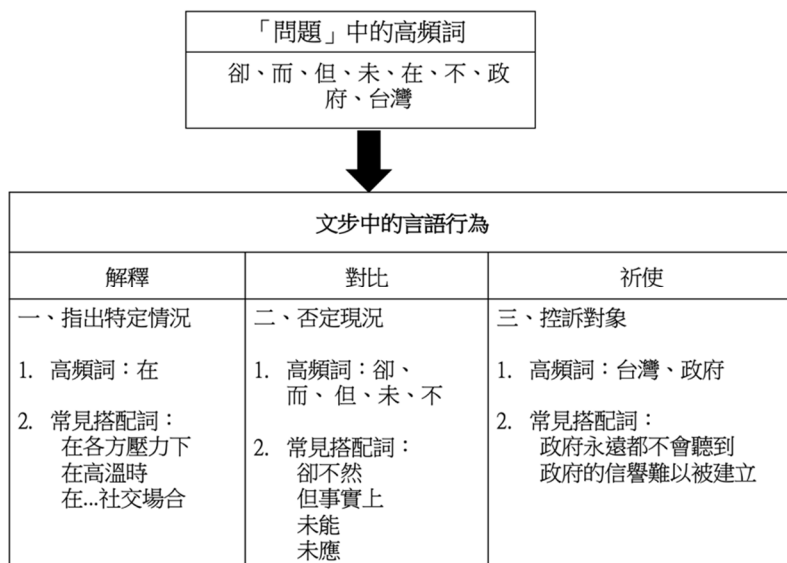


圖 7. 「問題」文步高頻詞與對應之言語行為
[Figure 7. High-frequency Words and the Corresponding Communicative Acts in the Problem Move]

4.3 解方文步 (Solution)

提案人在此文步中最常使用情態動詞「應」。鄭綦(2000)說明「應」與「應該」、「應當」、「該」往往被視為同義詞，且「應」似乎是較為文言的詞彙，「應該」或「該」屬口語的詞彙。因而可用來解釋為什麼語料中出現大量「應」，卻並未因此出現「應該」等同義詞。此外，根據鄭綦(2000)的觀察發現「應」在語料中的用法帶有義務的含義，書寫者認為讀者（政府）有義務去實施某件事，而這種「義務陳述」相當於 Blum-Kulka、House 和 Kasper (1989)所提出的直接請求策略之次要策略，是透過直接言語行為，直接地陳述請求。如下例(10)及(11)：

(10) 小型汽車置放架其使用應依下列規定。

(11) 中央政府應鼓勵設置社區型足球場。

「應」是出現頻率最高的動詞，而其餘動詞如「應該」、「應於」、「應有」則出現不超過 5 次。本研究也發現此文步中，提案人使用「有」、「以」加強立論基礎、鞏固論點，如例(12)。「解方」文步不僅蘊含請求以及鞏固論點的語言功能，也提出解決方法的可行性作為評估標準，具體說明該解方的優缺點，因此使用「能」表達意願與可能性，如例(13)。

- (12) 立刻公開台灣各地即時地磁訊號圖。人民**有**知的權利。
- (13) 以期將來**能**達到節省全民健康保險醫療費用之支出。

此外，研究者也發現「以」在解方文步中出現的次數偏多，「以」是一個較文言的書面語介詞，是中文裡特別的文法標記。相較於其他文類中使用「解決方法」、「解方」等相關抽象名詞帶出解方，此文類的寫作者更常使用「以」引介出解決方法。

- (14) 動保法中應明定條文，將保護動物理念納入我國教育體系的課綱中，以實現基本生命教育學程。

研究者在圖 8 整理出此文步中的言語行為與關鍵字。除了言語行外之外，本研究也發現運用條列式寫法、簡潔有力的文字表達訴求是「解方」文步中常見的現象，與其他文步相較之下，提案人在書寫「解方」時更偏向使用條列式整理出重點，以直接的語氣表達請求，因此在詞數方面，容易比其他文步少。

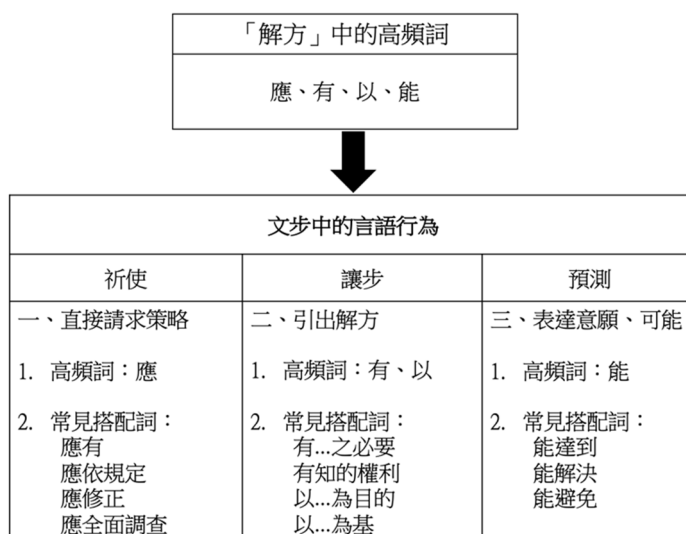


圖 8. 「解方」文步高頻詞與對應之言語行為
[Figure 8. High-frequency Words and the Corresponding Communicative Acts in the Solution Move]

4.4 評價文步 (Evaluation)

「評價」文步則對提案作出總結，評價從中獲得效益或是不足之處，作為未來修正的方向，期許將來能提出更好的解方。從表 1 的數據會發現，「評價」是必要的文步。從語料中發現，多數提案者傾向在此文步重述並總結前述所提及的「情況」、「問題」以及

「解方」，因此出現不少重複的描述，最後做出整體的「評價」；有些提案者則直接說明「解方」所帶來的效益，以及未來展望。

從語料中歸納此文步中最主要的兩個言語行為分別為：介進有關的事物、對象與提出正面的回應，使用「對」介進有關的事物、對象，舉出與事件有關的利害關係人，如例（15）。最後，使用「讓」、「更」、「能」做出正面回覆，有助於提出評價和未來願景，如例（16）。總結此文步的高頻詞和言語行為，研究者將之整理於圖9。

- (15) **對**酒駕犯罪者施以懲罰...；**對**其他社會大眾亦具有教育意義。
- (16) **也****能**為下一代帶來**更**好的教育與省思，謝謝大家的幫忙！

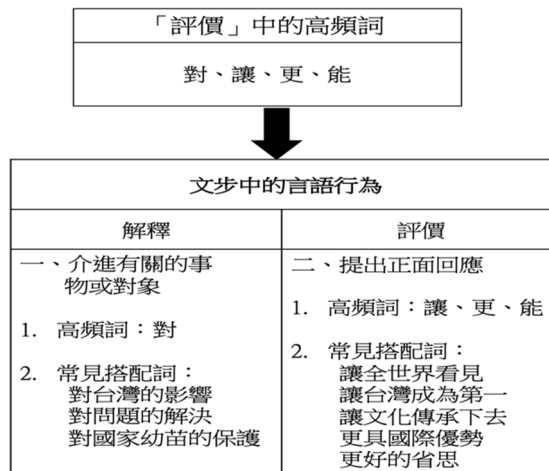


圖9. 「評價」文步高頻詞與對應之言語行為
[Figure 9. High-frequency Words and the Corresponding Communicative Acts in the Evaluation Move]

5. 結論 (Conclusion)

本研究分析「問題-解方」架構如何應用於網路提案型文章，找出必須出現與可有可無的文步，並且從各文步中的高頻詞排序中，找出能代表各文步的搭配詞和言語行為。研究發現，儘管在不同的文步中會出現不同的高頻詞如：情態動詞、介詞...等，而這些言語行為與高頻詞之間的關係並非是一對一的，如：「能」同時出現在「解方」和「評價」文步，在「解方」文步時，「能」的言語行為是「預測」，用來表達意願與未來事件的可能性；在「評價」文步時，「能」的言語行為是「評價」，用來提出正面回應。

本研究成果不僅回應了之前的文獻，也為中文文類分析拓展出新的研究領域，透過研究不同類型的文本，有助於探究出語言的不同意義。由於過去中文文類分析的相關研究僅限於學術領域，尚未出現分析網路論壇文章的相關研究，本研究首先從應用語言學的觀點切入，建立分析架構，試圖透過文步分析更加了解網路提案型文章的文類特色，

以及此文類中所含的文章架構、文步分佈以及詞彙特徵為何。透過了解不同文類中的詞彙特徵有助於找出詞彙變體，可作為模板開發文步分類器，朝向自動化分類的研究邁進。然而本文的研究限制為資料數量不足，因此缺乏更深、更廣地分析與討論。

表 6. 提案類型文本所涵蓋之文步、言語行為與高頻詞總整理
[Table 6. Communicative Acts and High-frequency Words in Each Move]

文步	言語行為	功能	高頻詞
情況	陳述	一、判斷某種情況	是、為
		二、指出時間點	年、在
		三、列舉相關人、事件	在、者、已
	凸顯	四、對現況提出否動或質疑	為、不、無法、卻
問題	解釋	一、指出特定情況	在
	對比	二、否定現況	卻、而、但、未、不
	祈使	三、控訴對象	台灣、政府
解方	祈使	一、直接請求策略	應
	讓步	二、鞏固論點	有、以
	預測	三、表達意願、可能	能
評價	解釋	一、介進有關的事物或對象	對
	評價	二、提出正面回應	讓、更、能

致謝 (Acknowledgements)

本論文承蒙科技部計畫(109-2410-H-004-163-)「基於語料庫之跨語言隱喻信號研究」及國立政治大學英國語文學系高教深耕計畫(110H123-06)「空殼名詞於跨語言之研究」的補助，以及兩位匿名審查委員的修改建議，特此致謝。

參考文獻 (References)

Austin, J. L. (1962). *How to do things with words*. Oxford university press.

- Ali, A. M. (2013). Combining problem-solution categories and communicative acts: An analysis of Malaysian and British business journalistic texts. *World Applied Sciences Journal*, 21, 174-185. <https://doi.org/10.5829/idosi.wasj.2013.21.sltl.2152>
- Belmonte, M. I. A. (2009). Positioning the reader: A study on the use of interactive textual patterns in English written newspaper editorials and articles of opinion. *English Text Construction*, 2(1), 48-69. <https://doi.org/10.1075/etc.2.1.03alo>
- Bhatia, V. K. (1993). *Analysis Genre: Language Use in Professional Settings*. Longman.
- Blum-Kulka, S., House, J., & Kasper, G. (1989). Cross-cultural pragmatics: Requests and apologies. *Grazer Linguistische Studien*, (Heft), 349-357.
- Charles, M. (2011). Adverbials of result: Phraseology and functions in the Problem-Solution pattern. *Journal of English for Academic Purposes*, 10(1), 47-60. <https://doi.org/10.1016/j.jeap.2011.01.002>
- Flowerdew, L. (2003). A Combined Corpus and Systemic-Functional Analysis of the Problem-Solution Pattern in a Student and Professional Corpus of Technical Writing. *TESOL Quarterly*, 37(3), 489-511. <https://doi.org/10.2307/3588401>
- Flowerdew, L. (2008). *Corpus-based Analyses of the Problem-Solution Pattern: A Phraseological Approach*. John Benjamins.
- Hagen, L., Harrison, T. M., Uzuner, Ö., May, W., Fake, T., & Katragadda, S. (2016). E-petition popularity: Do linguistic and semantic factors matter? *Government Information Quarterly*, 33(4), 783-795. <https://doi.org/10.1016/j.giq.2016.07.006>
- Heffernan, K., & Teufel, S. (2018). Identifying problems and solutions in scientific text. *Scientometrics*, 116(2), 1367-1382. <https://doi.org/10.1007/s11192-018-2718-6>
- Hoey, M. (1983). *On the Surface of Discourse*. Allen and Unwin.
- Hoey, M. (2001). *Textual Interaction*. Routledge.
- Hyland, K. (1990). A genre description of the argumentative essay. *RELC Journal*, 21(1), 66-78. <https://doi.org/10.1177/003368829002100105>
- Samraj, B. (2002). Introductions in research articles: Variations across disciplines. *English for Specific Purposes*, 21(1), 1-17. [https://doi.org/10.1016/S0889-4906\(00\)00023-5](https://doi.org/10.1016/S0889-4906(00)00023-5)
- Ratanakul, S. (2017). A study of problem-solution discourse: Examining TED talks through the lens of move analysis. *LEARN Journal: Language Education and Acquisition Research Network*, 10(2), 25-46.
- Ratanakul, S. (2018). A Move Analysis of Problem-Solution Discourse: A Pedagogical Guide for Opinion and Academic Writing. *Arab World English Journal*, 9(3), 233-247. <https://doi.org/10.24093/awej/vol9no3.16>
- Searle, J. R. (1976). A classification of illocutionary acts. *Language in Society*, 5(1), 1-23.
- Swales, J. M., (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press.
- Swales, J. M., & Feak, C. B. (2012). *Academic Writing for Graduate Students: Essential Tasks and Skills*. University of Michigan Press.

- Vestergaard, T. (2000). *From genre to sentence: The leading article and its linguistic realization*. In F. Ungerer (ed), *English Media Texts – Past and Present: Language and Textual Structure*, 151-176. John Benjamins.
- 鄭縈 (2000)。從語料庫看漢語助動詞的語法特點。In *Proceedings of Research on Computational Linguistics Conference XIII(ROCLING XIII)*, 157-170。[Cheng, Y. (2000). The Syntactic Characteristics of Chinese Auxiliaries based on the Sinica Corpus. In *Proceedings of Research on Computational Linguistics Conference XIII(ROCLING XIII)*, 157-170.]
- 何淑貞、王錦慧 (2010)。華語教學語法。文鶴出版。[Ho, S.-C. & Wang, J.-H. (2010). *Chinese Pedagogical Grammar*. Crane.]
- 黃冠誠、吳鑑城、許湘翎、顏孜曦、張俊盛 (2014)。學術論文簡介的自動文步分析與寫作提示。中文計算語言學期刊, 19(4), 19-46。[Huang, G.-C., Wu, J.-C., Hsu, H.-L., Yen, T.-H., & Chang, J. S. (2014). Automatic Move Analysis of Research Articles for Assisting Writing. *International Journal of Computational Linguistics & Chinese Language Processing*, 19(4), 19-46.]
- 陳坤毅 (2019)。影響政府回應的提案因素—以公共政策網路參與平臺為例。國立臺灣大學社會科學院公共事務研究所碩士論文。[Chen, K.-y. (2019). *The Determinants of E-petitions on Government Responsiveness: A Case Study of the Join Platform*. M.A. Thesis, National Taiwan University.]

An N-gram Approach to Identifying the Chinese Linguistic Signals for the Problem-Solution Pattern in Annotated Online Health News

Chen-Yu Chester Hsieh* and Yu-Yun Chang⁺

Abstract

This article will report the results of an exploratory project that combined the annotation of the Problem-Solution (PS) textual pattern in online health news and the quantitative and qualitative methods of corpus linguistics to investigate the linguistic features of particular rhetorical moves. A total of 120 journalistic texts written in Chinese were collected from a Taiwan-based journalistic website that focused on providing news related to health and medicine and were annotated with the four components of the PS pattern. To identify signals in the genre for the elements of the PS move structure, an n-gram approach was then implemented to extract frequent lexicogrammatical sequences from the corpus in general and from the Problem and Response moves in particular. The results showed that the linguistic features found in the retrieved sequences tended to fall within a range of categories, such as abstract nouns, medical terms, and modal verbs, which not only served as functions relevant to the rhetorical move in which they were used but also reflected characteristics specific to the health news genre and the Chinese language. The findings and annotated data generated from the current project will thus provide a solid foundation for future research and applications.

Keywords: N-gram, Problem-Solution Pattern, Health News, Annotation, Journalistic Discourse

* Research Center for Translation, Compilation and Language Education, National Academy for Educational Research, New Taipei City, Taiwan

E-mail: chesterhugues@gmail.com

⁺ Graduate Institute of Linguistics, National Chengchi University, Taipei City, Taiwan

E-mail: yuyun@nccu.edu.tw

The author for correspondence is Yu-Yun Chang.

1. Introduction

As problem solving constitutes an essential activity and ability in various fields (Jonassen, 2000; Charles, 2011; Handford & Matous, 2015; Heffernan & Teufel, 2018), the Problem-Solution (PS) pattern is arguably one of the most common rhetorical structures across cultures and text types (e.g., Hoey, 2001; Flowerdew, 2003; Charles, 2011; Ratanakul, 2017). According to Hoey (1983, 2001), the PS pattern (also known as the SPRE [Situation-Problem-Response-Evaluation] or SPSE [Situation-Problem-Solution-Evaluation] pattern) is typically composed of four moves: Situation (i.e., background information regarding the topic or problem), Problem (i.e., issues that need to be addressed or fixed), Response/Solution (i.e., proposals or methods for dealing with the problem), and Evaluation (i.e., assessments of the mentioned solutions/responses).

Due to its prominence and popularity in English texts, the PS rhetorical pattern has garnered the attention of several text/corpus linguists (e.g., Hoey, 1983, 2001; Scott, 2000; Flowerdew, 2003, 2008; Charles, 2011). Previous findings have revealed that the PS pattern appears in a wide range of genres, such as advertisements (Hoey, 2001), academic reports (Flowerdew, 2003), research articles (Charles, 2011; Heffernan & Teufel, 2018), journalistic texts (Ali, 2013; Ratanakul, 2018), and even TED talks, (Ratanakul, 2017). The moves of the pattern may be ordered in different sequences, with some of the components repeated or excluded (Hoey, 2001; Belmonte, 2009; Ratanakul, 2018), and they are often signaled by particular identifiable linguistic features, such as abstract nouns (Flowerdew, 2003, 2008) and adverbials (Charles, 2011), and are supported by various communicative functions (Belmonte, 2009; Ali, 2013).

Despite the number of studies on this topic, a few research gaps still remain. First, most, if not all, of the past studies drew on data written or delivered in English. However, no evidence indicates that this discourse pattern is restricted to the English language. Recent research has suggested that this rhetorical pattern is prevalent in English texts produced both inside and outside the Inner Circle (Kachru, 1992), where English is learned and used as the first or native language (e.g., Ali, 2013; Handford & Matous, 2015; Ratanakul, 2018). Scholars such as Belmonte (2009) and Ali (2013) have also advocated the compilation of multilingual corpora or a cross-linguistic comparison of the PS pattern. Applying this textual structure in the analysis of articles composed in Chinese would allow for a closer look at how this textual pattern is signaled and realized in typologically and culturally distinctive languages.

Another major gap in the literature pertains to the genres that have been investigated. Although journalistic writing is one of the most often studied text types with respect to the SPRE pattern, only business news texts and opinion articles have been examined in more detail (e.g., Belmonte, 2009; Ali, 2013). Medical and health news, which presents information pertinent to medical conditions and their treatments (李松濤 & 鄔啟柔, 2017; 李松濤 & 許

文怡, 2020), has not yet been scrutinized under the framework of the PS pattern. As has been noted by scholars, journalistic texts regarding medical and health issues have substantial influence over readers' well-being and lifestyle decision-making (Fu *et al.*, 2020; De Coninck *et al.*, 2020), and may even shape their news and scientific literacies (李松濤 & 鄔啟柔, 2017; 李松濤 & 許文怡, 2020); as such, this particular genre is worthy of further investigation.

Finally, most of the previous studies on the PS textual structure adopted a keyword analysis approach (Scott, 2000; Flowerdew, 2003, 2008) or simply focused on a given part of speech or communicative act (Charles, 2011; Ali, 2013). Although these methods may be suitable for English, a language that has clearer word boundaries, less disputable syntactic categories, and more literature related to the topic in question, they may not be as applicable to analyzing texts written in languages that do not have the above characteristics, such as Chinese (Fu *et al.*, 2020; Tian *et al.*, 2020). As a result, an even more data-driven approach, such as n-gram analysis, may be preferable for an exploratory investigation of the linguistic signals for an understudied rhetorical pattern in articles composed in Chinese.

In light of the above gaps, the current paper aims to report the results of a preliminary project that examined the PS pattern in online health-related journalistic articles from one of the most popular and influential health news websites based in Taiwan by integrating the annotation of the PS discourse structure and the methods of corpus linguistics. In addition to not having been investigated under the SPRE framework, the genre of online health news was chosen because it has become one of the major sources of medical information for the general public, with real-life consequences (Fox & Duggan, 2013), and because gaining more knowledge about how this genre is composed and organized will allow for opportunities to improve media and science literacies (李松濤 & 鄔啟柔, 2017; 李松濤 & 許文怡, 2020) and to develop algorithms that can help detect and retrieve information about particular medical conditions and recommended solutions, facilitating the process of solving health-related problems (cf. Ghenai & Mejova, 2018; Waszak *et al.*, 2018; Kumari *et al.*, 2021; Tsirintani, 2021). More specifically, the current study set out to answer the following research questions:

- RQ 1: What are the distribution patterns of the SPRE moves in online health news articles written in Chinese?
- RQ 2: What are the most common Chinese n-grams (trigrams in this study) in the online health news articles and what features do they display?
- RQ 3: How do the most frequent trigrams reflect the characteristics of this genre?
- RQ 4: What are the most frequent and most distinctive trigrams in the Problem and Response moves of the online health news articles? What are some of their generalizable features?
- RQ 5: How do these trigrams help achieve the rhetorical functions of the Problem and Response

moves in the online health news articles?

By answering the above questions, this study demonstrated that the n-gram approach was a useful framework for retrieving the linguistic signals of specific move components, the results of which will add to prior research on move structures, journalistic genres, and the annotation of Chinese texts. Moreover, by using a small set of data, the findings of the current study, which was exploratory in nature, will serve as the foundation for future research using a larger amount and greater diversity of data.

2. Related Works

As one of the earlier scholars that paid close attention to the PS pattern in English, Hoey (2001) put forth a number of insightful observations about this discourse structure. First, the PS pattern is created by the writer of a text to respond to a series of questions although the order of the moves is not fixed. Second, the linguistic signals in the PS pattern are typically lexical items, including words that are explicitly related to the moves, such as “solution” and “problem,” and evaluative devices, such as “unfortunately” and “have no money.” Third, the Situation move is often retrospectively identified and an intervening stage, such as planning or recommendation, can be found between the Problem move and the Solution move. Lastly, the PS pattern can be recycled when more than one solution is presented. Therefore, in lieu of the Situation-Problem-Response-Evaluation pattern, the move structure can be as complex as Situation → Problem → Response 1 → positive Evaluation → negative Evaluation → Response 2 → positive Evaluation (Hoey, 2001: 134).

2.1 The PS Pattern in Academic Writing

Based on Hoey’s (2001) insights, a number of researchers have investigated the realization of the PS pattern in various contexts. One of the most investigated genres in the literature is academic writing (Flowerdew, 2003; Charles, 2011; Heffernan & Teufel, 2018). For example, conducting a keyword analysis to compare the PS pattern in the technical academic reports composed by novice and professional writers, Flowerdew (2003; 2008) observed that the Problem move was closely linked to causal relationships, such as Reason-Result and Means-Purpose, and was often marked by lexical devices, including connectives such as “therefore” and “as a result,” causative verbs such as “lead to” and “avoid,” and abstract nouns such as “problems” and “impacts.” Moreover, the nominalization forms of verbs such as “implementation” and “recommendation” together with verbs in the present perfect such as “recommended” were found to be indicative of the Solution element in reports composed by professional writers (Flowerdew, 2008).

With a focus on the adverbials of “result” and “contrast” in theses written by native-speaker writers in politics and in materials science, Charles (2011) analyzed the functions and co-occurrence patterns of the most frequent adverbials of “result” (i.e., “thus”) and “contrast” (i.e., “however”) in the PS pattern and found that despite the differences in number between disciplines, in general, “however” was used to mark the Problem element and “thus” was used to signal the Evaluation of the Response move when the former preceded the latter. On the other hand, when the order was reversed, while “however” was still the signal for the Problem move, the preceding “thus” marked the Situation move instead.

More recently, Heffernan and Teufel (2018), drawing on a machine learning approach and a corpus of problem and solution statements, specified 15 features to help develop classifiers that could identify Problem and Solution elements in scientific texts, including n-grams, polarity, and syntax. Heffernan and Teufel (2018) asserted that their model accurately differentiated problems/solutions from non-problems/solutions and that syntactic information, documenting, and word embedding were the three best features that allowed them to achieve the target task.

2.2 The PS Pattern in Journalistic Texts

Another major line of research looked into how the PS pattern was realized in various types of journalistic genres, including feature articles (Scott, 2000), opinion pieces (Belmonte, 2009; Ratanakul, 2018), and business news (Ali, 2013). Scott (2000), one of the earliest to examine the use of the PS pattern in journalist texts, found that, surprisingly, the nouns “problem” and “solution” were not especially frequent and thus not “key” enough in the Guardian’s feature articles. Even when the two nouns were characterized as keywords, their signaling function tended to be local rather than global, in contrast to what had been commonly presumed.

Belmonte (2009), investigating the rhetorical organization of editorials and op-eds in USA Today, identified a number of characteristics specific to those genres with regard to the PS textual pattern. First, the textual pattern in the two genres tended to revolve around the Problem and Evaluation elements, while the Solution move was not as prominent. This indicated that it was a genre convention for the editorial writer to leave the solution to the problem unspecified. Second, different components of the PS structure displayed the tendency to occur in different parts of the articles, each with a particular rhetorical function. For example, when the Evaluation element was presented at the end of a text, it was usually more negative, creating a feeling of discomfort or dissatisfaction. Third, despite the fact that editorials and op-eds are both, in general, evaluative texts supported by other rhetorical roles, such as elaboration and justification, and by other communicative acts, such as statements and assertions, the PS pattern in editorials tended to show a more impersonal style, while that in op-eds was more often presented in a more involved style, with more frequent use of shared knowledge assertions and recommendations.

Similarly, Ratanakul (2018) carried out a move analysis of opinion columns in two newspapers online, the New York Times and China Daily, based on the SPRE framework. Partially echoing Belmonte's (2009) findings, Ratanakul (2018) also noted that the Problem move constituted the most central component of the PS pattern in the articles, although the Response element was found to be the second most frequent move in the data, in contrast to what Belmonte (2009) reported. In addition, Ratanakul (2018) pinpointed several features of each move in the PS pattern, both obligatory and optional, including causes of the problem (Problem) and a call for action (Solution), which indicated that each component of the target discourse structure could be further divided and analyzed.

On the other hand, Ali (2013) examined the PS textual pattern and its communicative functions, such as informative, evaluative, explicative, and predictive, found in journalistic articles in business magazines published in Malaysia (i.e., Malaysian Business [MB]) and the United Kingdom (i.e., Management Today [MT]), respectively. In agreement with previous studies such as Belmonte (2009), Ali's (2013) findings suggested that elements in the PS pattern, such as Problem and Solution, contained multiple speech and communicative acts and that the frequency and distribution of each act in the moves differed across sources of the texts. Nevertheless, no one-to-one connections were found between the communicative acts and the Problem and Solution components in the business news articles. For example, informative, evaluative, and predictive acts were identified in both the Problem and Solution moves.

All in all, the works reviewed above point to the fact that while the PS pattern is a prevalent rhetorical structure across genres, factors such as topics, text types, and language variations influence the realization of the PS pattern. An analysis of this pattern in languages other than English, as several researchers have suggested, is thus warranted. The present study aimed to contribute to this line of research.

2.3 Functional Classification of N-grams

In the realm of applied linguistics, n-grams are studied under different labels, such as "lexical bundles" (Biber *et al.*, 2004) and "multiword expressions/sequences" (e.g., Staples *et al.*, 2013). One of the most widely adopted functional taxonomies was proposed by Biber *et al.* (2004), in which lexical bundles serve as the functions of "stance," "referring," and "discourse organizing." Stance bundles allow users to mark an epistemic or attitudinal modality (e.g., 'I don't know if' and 'if you want to'), while referring bundles are used to identify an entity, convey imprecision, specify an attribute, or refer to a particular place, time, or part of the text (e.g., 'that's one of the', 'something like that', 'there's a lot of', and 'at the same time'). Finally, discourse organizing bundles introduce a focus or manage topics (e.g., 'take a look at' and 'on the other hand').

While Biber *et al.*'s (2004) framework is the most general, commonly used functional

taxonomy, a number of more genre-specific classifications have also been proposed. For example, Biber and Gray (2013) modified the taxonomy put forth by Biber *et al.* (2004) for the analysis of responses in the TOEFL iBT speaking and writing tests. To accommodate the purpose of the writing tasks, they divided the stance bundles into two subgroups, “personal/epistemic” and “attitudinal/evaluative,” and the discourse organizing bundles into three subtypes, “information source,” “information organizer” (marking more specific information in the writing), and “discourse organizer” (serving more general discourse functions in the discourse), expanding the taxonomy into a five-way categorization (notably, this classification does not include the referring function). Biber and Gray (2013) found that the distribution pattern varied in accordance with the task type and the writer’s performance level.

To investigate research articles, master’s theses, and dissertations from different disciplines, Hyland (2008) developed a functional taxonomy that grouped n-gram expressions according to whether they were “research-oriented,” “text-oriented,” or “participant-oriented.” In that taxonomy, research-oriented bundles refer to expressions that facilitate writers’ presentation of experiences and activities in the real world, such as location and procedure, whereas text-oriented bundles pertain to the organization of texts, such as transition signals. Finally, participant-oriented bundles are sequences of words that involve the writer and/or the reader of the text and may serve functions related to stance and engagement (cf. Hyland, 2008).

On the other hand, some researchers categorized n-gram expressions based on their function in rhetorical moves. For example, analyzing the sentence-initial multiword expressions in Arts and Humanities PhD dissertation abstracts, Li *et al.*’s (2020) tailor-made functional taxonomy consists of “background bundles,” “purpose bundles,” “method bundles,” “findings bundles,” “implications bundles,” and “structure bundles.” According to Li *et al.* (2020), each of these types helped achieve the rhetorical function of a specific move in the abstracts, such as stating the research purpose or outlining the structure.

3. Methodology

3.1 Data Collection and Annotation

We targeted the health topic for the analysis of the moves in the PS discourse structure since it has been widely abused and propagated through content farm articles on the Internet. Several studies have worked on analyzing and detecting the spread of health misinformation (Ghenai & Mejova, 2018; Waszak *et al.*, 2018; Kumari *et al.*, 2021; Tsirintani, 2021). An increasing number of people prefer to seek health and medical solutions by themselves first rather than directly consult medical professionals. The Pew Research center conducted a survey in 2013 and found that 59% of adults sought health information online and 3% were harmed by misinformation (Fox & Duggan, 2013).

To build a corpus annotated with the moves in the PS pattern, articles published on the Heho 健康 website were retrieved and collected.¹ Heho is a popular health-related website in Taiwan that provides credible health-related information contributed and certified by healthcare professionals. Although there are other platforms that publish and disseminate health information to the public, the content is often not provided by reliable sources and may have been released by content farms. Since many people prefer to seek solutions for health issues online beforehand, we focused on investigating the PS discourse structure for content that involved questions frequently asked by people and solutions given by medical experts. To the best of our knowledge, no research has explored the online health information issue under the scope of the PS discourse structure. Thus, 120 articles (139,131 tokens) published by 10 journalists in the series 醫生說 ‘The Doctor Says’ under the 請問專家 ‘Asking Experts’ topic were collected, with the publication dates ranging from March 11 to July 12, 2021. The retrieved articles in the series drew on various expert sources and involved specific health-related problems and solutions. Another characteristic of the articles was that the main problem highlighted by the authors was addressed directly in the articles’ titles.

To annotate the articles, we adopted the definitions of the PS moves specified in Ratanakul (2018: 235) to identify each element of the PS pattern, as listed below:

- Situation: background information on situations; facts about people, issues, events, places involved in the issue of discussion
- Problem: aspect of a situation requiring a response, need, dilemma, puzzle, or obstacle under discussion; weaknesses inherent to the current situation
- Response: solution(s) to the problem; discussion of a way(s) to deal [with] or to solve the problem
- Evaluation: assessment of the effectiveness of the proposed solution(s); if there is more than one solution, which solution is the best?

To increase the efficiency of the annotation task, we utilized one of the prevalently adopted annotation tools in corpus linguistics, GATE.² Although the definition of the annotation scheme was provided, most previous studies that annotated the PS pattern simply made up or took some examples from the texts for analysis, without annotating the entire article, because the annotation scheme as well as the PS pattern were proposed at the discourse level instead of the sentence level, which made it difficult to set clear boundaries for annotation. In this study, we followed the definition of the annotation scheme and annotated the PS pattern throughout all the articles. As each of the moves could occur more than once in an article, several moves

¹ The Heho 健康 website is available at: <https://heho.com.tw/>

² The GATE annotation tool is available at: <https://gate.ac.uk/download/>

in the same article were annotated with the same move label. Thus, each article was labeled with several moves of Problem, Response, Situation, and Evaluation. In the dataset, since the titles of the articles revealed the main problems to be discussed, each annotator was asked to first identify the problem(s) in an article based on its title. The title of one of the retrieved articles is shown in (1). As suggested by the title, the main problem discussed in this article was 剖腹產後會有沾黏風險 ‘there is a risk of adhesion after a C- section’. Based on the identified problem, the PS moves in the article were then annotated accordingly.

- (1) 剖腹產後會有沾黏風險！產科權威提醒「抗沾黏」3 大步驟
‘There is a risk of adhesion after a C-section! The obstetrician authority suggests three steps for anti-adhesion’

To ensure that the annotators followed and understood the annotation scheme consistently, a training session with labeled articles was held before the annotation task was performed. In order to maintain annotation quality during the task, a checkpoint meeting was arranged. In the meeting, all the annotation issues raised by the annotators were discussed. The annotators were asked to modify their annotations based on the discussions. Since it would take more time for the annotators to label the PS pattern throughout the entire article, each article was annotated by one annotator. Four annotators were recruited for this annotation task. All 120 articles collected were equally distributed to the four annotators, so each annotator was assigned 30 different articles.

3.2 Data Analysis

In this study, the trigrams were character-based instead of word-based under the following considerations. Character-based n-grams have been widely applied in different tasks, such as text classification (Cavnar & Trenkle, 1994), spam filtering (Kanaris *et al.*, 2007), authorship attribution (Escalante *et al.*, 2011), and plagiarism detection (Kuta & Kitowski, 2014). Some studies have demonstrated that character-based n-grams are more effective in generating word embeddings for unknown words (Wieting *et al.*, 2016; Bojanowski *et al.*, 2017) and are more informative in performing topic categorization and document summary (Giannakopoulos & Karkaletsis, 2009). Since Chinese is not a space delimited language, the issue of defining a Chinese word has long been discussed and is still disputable (Ng & Low, 2004; Tian *et al.*, 2020). Although several Chinese segmenters have been released, segmentation results are still hard to evaluate and segmentation errors have led to coarse-grained data analysis. Moreover, based on the limited number of annotated articles, processing word-level n-grams in a dataset has also led to fewer n-gram types, resulting in less statistical effectiveness in analyzing PS

patterns.³ Therefore, we considered each Chinese character the basic unit of analysis in this study and targeted trigrams for further exploration. We also applied the Natural Language Toolkit in Python to retrieve trigrams from the collected dataset.⁴

N-gram patterns with a larger n may occur only once in a dataset, resulting in relative frequencies that are close to zero. It is difficult to observe the PS pattern based on raw frequencies and relative frequencies, since both values cannot be placed on a normalized scale for exploration. Different scaling approaches (e.g., z-score and Min-Max) may be applied to help normalize the dataset before investigation. The normalization techniques of z-score and Min-Max scaling are different. Z-score scaling takes the entire distribution of a dataset into consideration by calculating the mean and standard deviation, and the range of the z-score is affected by the dataset distribution accordingly, either positive or negative. Min-Max scaling has the benefit of scaling all the values into positives and transforms the values into a range between 0 and 1. Since it is easier to inspect a dataset with a fixed range of normalized values, we chose to apply Min-Max scaling in this study.

After the target trigrams were retrieved and selected, the trigram expressions most prevalent in the two core moves of the PS pattern (i.e., Problem and Response) were identified and treated as signals for the respective moves. To analyze how these signals helped achieve the communicative goals of the moves, the trigrams were manually checked and categorized structurally according to the part of speech and semantic features of the main component and functionally based on Biber *et al.*'s (2004) tripartite taxonomy. The functional categories consisted of (i) stance (including epistemic and attitudinal stances); (ii) discourse organizing (including marking the information sources, managing the topic and focus, and signaling other

³ We performed word-level trigrams with the jieba segmenter in our preliminary study, as suggested by one of the reviewers. After segmentation, the highest frequency of a word-level trigram in the corpus was 22, which was significantly lower than the highest frequency of a character-level trigram (i.e., 113). The top four word-level trigrams were 異位性皮膚炎 'atopic dermatitis', 機器手臂 'robotic arms', 李孝貞醫師說 'the doctor Xiao-Zhen Li says', and 的情況下 'situation of'. Compared with the character-level trigram results listed in Table 2, the word-level trigrams were mostly technical or topic-specific terms, rather than carrying the linguistic cues of the PS pattern. Thus, character-level trigrams were taken into consideration in exploring the PS pattern in this study.

⁴ In the present study, trigram sequences in particular were examined for further analyses for a number of reasons. First, a number of bigram sequences did not convey a complete meaning (e.g., ◦ 丿 and 丿 如) and thus did not produce enough PS patterns for analysis. Second, the number of sequences was too high for more qualitative analysis. Third, sequences that included more than three characters were either specific technical terms instead of patterns, such as 攝護腺癌 'prostate cancer' (4-gram), 代謝症候群 'metabolic syndrome' (5-gram), and 異位性皮膚炎 'atopic dermatitis' (6-gram), or too few in frequency for generalization. As a result, trigram sequences were the best unit of analysis for the current research.

discourse relations); and (iii) referring (including identifying the focus of the move and referring to circumstantial elements such as time and space).

4. Results and Discussion

4.1 Descriptive Statistics of the Annotated Dataset

Among the 120 articles collected, 113 articles were judged as containing the PS pattern (henceforth target articles), and the remaining seven articles were ignored and left unannotated since they were not the focus of this research. The structure of each article contained an average of 13 paragraphs (SD=4.45) and 1,159 tokens (SD=370), and 1,212 items were identified and annotated based on the four moves.

Table 1. Number of items in each of the four PS moves in the target articles

Moves	Situation	Problem	Response	Evaluation
Number of Items	156	357	672	27

Table 1 above presents the number of moves observed in the target articles. As can be seen, among the components of the PS pattern, the Response move and Problem move were substantially more addressed and employed than the other two moves (e.g., Situation and Evaluation) (cf. Scott, 2000; Hoey, 2001). This tendency may have been partly due to the characteristics of online health news (i.e., readers want to see the relevance of the problem and identify the recommended solution(s) within a short time, and they do not prefer to read too much technical information [cf. 葉蓉慧 & 黃曄超, 2017]).

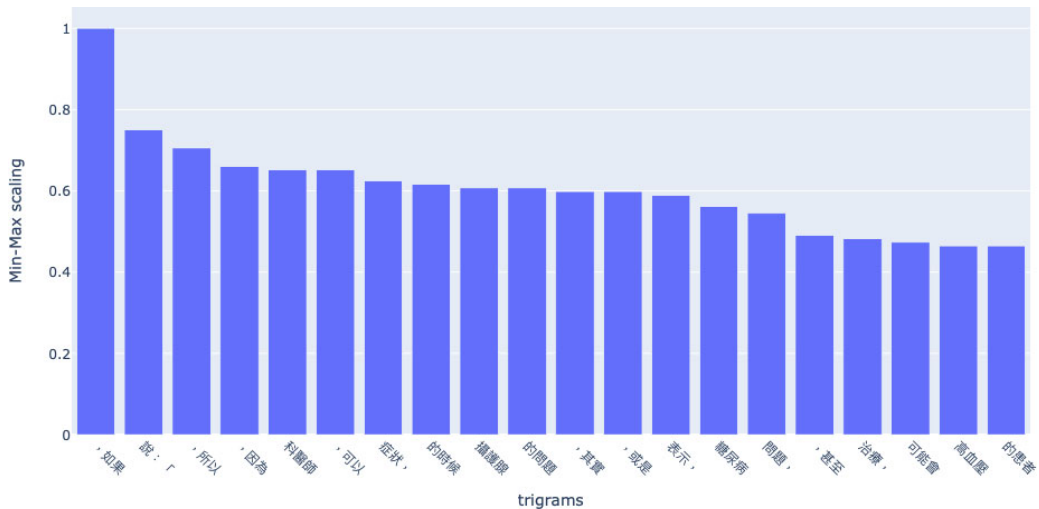
To further investigate the linguistic features of the PS pattern in a corpus-driven and quantitative way, we extracted trigrams from the corpus. Since the raw frequency of n-grams can be biased by the corpus size, all the trigrams were further normalized by Min-Max scaling. Table 2 below presents the top 10 frequently occurring trigrams with their corresponding statistical scores:

Table 2. Top 10 frequently occurring trigrams

Trigrams	Frequency	Relative Frequency	Min-Max
，如果	113	0.0009	1.0000
說：「	85	0.0007	0.7500
，所以)	80	0.0006	0.7054
，因為	75	0.0006	0.6607
科醫師	74	0.0006	0.6518
，可以	74	0.0006	0.6518
症狀，	71	0.0006	0.6250
的時候	70	0.0006	0.6161
攝護腺	69	0.0005	0.6071
的問題	69	0.0005	0.6071

As shown in Table 2, the relative frequencies of the trigrams were rather small as the largest value was only 0.0009, and they were highly sensitive to the corpus distribution and thus were not suitable to be used for observing patterns. After transforming the values via Min-Max scaling, the values were larger and thus easier to utilize for exploring the patterns.

Figure 1 shows an overview of the top 20 trigrams transformed via the Min-Max scaling approach:

**Figure 1. Min-Max scaling of the top 20 trigrams**

The threshold of the Min-Max scale value for the trigram expressions that would be further analyzed was set at 0.3 to avoid hapax legomenon. Since the dataset was small, a Min-Max

scaling threshold of 0.2 may have resulted in instances of hapax legomenon. The threshold of 0.3 allowed a total of 58 trigram sequence types to be selected. Items containing terms that were too technical or topic specific, such as 攝護腺 ‘prostate’, 糖尿病 ‘diabetes’, and 高血壓 ‘hypertension’, were eliminated from the list. Lastly, trigrams with a DP value higher than 0.85 based on the dispersion measure developed by Gries (2008) were excluded to avoid trigrams that occurred skewedly in a limited set of texts.⁵ As a result, 48 types of trigram items were included for later analysis, and the full list is displayed in Table 3. It should be noted that, as can be seen in the following table, trigrams that included punctuation marks were retained for further analysis. Although in English-based n-gram studies punctuation marks are usually not counted as an element of an n-gram because they are found to convey discourse relationships (Yue, 2006) and occupy the same amount of space as a character does in written traditional Chinese, they were not eliminated from the data. These trigrams were then manually examined for structural (part-of-speech) and functional analysis.

Table 3. Target trigrams for further analysis

，可以	，可能	，因此	，因為	，如果	，其實
，或是	，所以	，甚至	，通常	，像是	來說，
治療，	表示，	症狀，	問題，	疾病，	問題。
解釋：	說：「	釋：「	常見的	這樣的	的治療
的狀況	的風險	的原因	的時候	的疾病	的症狀
的方式	的問題	的患者	的情況	的傷口	治醫師
是因為	可能會	有可能	就可以	越來越	會出現
會造成	臨床上	醫師說	科主治	科醫師	主治醫

4.2 General Observations of the Frequent Trigram Expressions

Far from being highly diversified, the most frequent trigram sequences were restricted to merely a number of categories, including connectives, adverbials, abstract nouns, health- and medicine-related terms, quotative devices, and a few others. This suggested that the genre of online health news as well as the PS pattern in that genre tended to feature the use of specific lexicogrammatical signals. For example, echoing Flowerdew’s (2003, 2008) observations, linguistic devices that marked the PS pattern were predominantly indicative of the Reason-Result relationship, including 因為 ‘because’, 所以 ‘so’, 因此 ‘therefore’, 造成 ‘to

⁵ A DP value of a trigram is a number between 0 and 1. The higher the value, the more skewed the distribution of the trigram in the corpus is. See Gries (2008) for more detailed information about this dispersion measure.

cause’, and 的原因 ‘the reason of’. Similar to the findings of previous studies (e.g., Hoey, 2001; Flowerdew, 2003, 2008), abstract nouns that were semantically connected to components of the PS pattern, explicitly or implicitly, such as 問題 ‘problem’, 風險 ‘risk’, 狀況 ‘condition’, 方式 ‘method; approach’, 原因 ‘reason’, and 情況 ‘situation’, were also found in the most frequent sequences. More notably, most abstract nouns were preceded by the grammatical morpheme 的, which links abstract nouns to another lexical or clausal unit. The preceding morpheme 的 can also serve as (part of) a shell noun construction (Schmid, 2000) that links abstract nouns to a co-referring phrase or clause (cf. Hsieh, 2020).

Results not reported in previous studies were also retrieved from the health news corpus. First, the conditional marker 如果 ‘if’ constituted the most frequent trigram sequence in the dataset. As pointed out by Lin (2019), conditionals play a crucial role in both spoken and written medical communication, serving multiple functions such as presenting suggestions, explaining successive treatment, and calling for attention. Another group of lexical items that was also less often mentioned in the literature was modal verbs, including 可以 ‘can’ and 可能 ‘may’, which supported communicative acts such as making predictions and recommendations in this genre (Hsieh, 2005). Finally, in correspondence with Flowerdew’s (2008) findings, verbs that indicated causal relationships such as 造成 ‘to cause’ were also found in the most recurrent trigram sequences, although not as prevalent as Flowerdew (2008) suggested.

Finally, features that were heavily linked to a particular genre or topic were spotted in the list of trigram sequences as well. For example, several nouns in the realm of physiology and medicine were involved in many of the sequences, including 醫師 ‘doctor’, 症狀 ‘symptom’, and 傷口 ‘wound’. On the other hand, insofar as the data were journalistic texts in essence, reporting verbs such as 說 ‘say’, 表示 ‘indicate’, and 解釋 ‘explain’ (Liu & Chiang, 2008) and typographical markers for quotations such as colons and quotation marks, which introduced quotes from the sources, were also found to be highly frequent (cf. Waugh, 1995), a tendency that has, again, rarely been reported in prior research. As will be shown in the following, many of these topic- and genre-oriented features were also signals for a particular component in the PS textual pattern.

4.3 Trigram Sequences in Rhetorical Moves

This section will focus on the trigram sequences that most frequently occurred in the Problem and Response moves, respectively, because these two moves not only theoretically constituted the most essential components of the PS pattern but also empirically accounted for the most move instances and contained the most high-frequency trigram expressions in the corpus. As will be presented later, although the Problem and Response moves shared a few types of trigram sequences with regard to their parts of speech, the sequences most commonly found in each of the two moves in fact differed from each other semantically and functionally, reflecting the

purposes that each move was intended to achieve.

4.3.1 Signals for the Problem Move

To begin with, a majority of the frequent trigram expressions in the corpus introduced above most frequently occurred in, and thus could be considered signals for, the Problem move. This move contained the widest range of trigram signals, including abstract nouns, medical terms, adverbials, connectives, reporting verbs, causal verbs, modal verbs, and others, as displayed in the following Table 4:

Table 4. Structural categories of the trigram signals for the Problem move

Abstract Nouns	Medical Terms	Reporting Verbs	Modal Verbs
的時候 ‘time of’ 的問題 ‘problem of’ 的原因 ‘reason of’ 的狀況 ‘situation of’ 問題。 ‘problem.’ 的風險 ‘risk of’ 問題， ‘problem,’	科醫師 ‘doctor of the Division of’ 主治醫 ‘attending physician’ 治醫師 ‘attending physician’ 科主治 ‘attending physician at the Division of’ 症狀， ‘symptom,’ 疾病， ‘disease,’ 的疾病 ‘disease of’ 臨床上 ‘clinically’	說：「 ‘say’ 醫師說 ‘doctor says’ 表示， ‘indicate,’	可能會 ‘probably will’ ，可能 ‘, probably’
Other Verbs	Connectives	Adverbials	Others
會造成 ‘will cause’ 會出現 ‘will appear’	，所以 ‘, so’ 是因為 ‘be because’	，其實 ‘, actually’ ，甚至 ‘, even’ ，通常 ‘, usually’	常見的 ‘common’ ，像是 ‘, such as’ 來說， ‘as for,’ 這樣的 ‘such; this type of’

With regard to abstract nouns, in agreement with Hoey (2001) and Flowerdew (2003, 2008), the Problem move tended to be marked by nouns that denoted or implied a problem, such as 問題 ‘problem’ and 風險 ‘risk’, and nouns that indicated a Reason-Result relationship, such as 原因 ‘reason’, as illustrated in (2) and (3), respectively. As mentioned earlier, these abstract nouns were mostly preceded by the functional morpheme 的 in the trigram sequences, which linked the abstract nouns to their co-referring components. In addition, nouns that referred to a circumstance, such as 狀況 ‘situation’ and 時候 ‘time’, were also found to be frequent trigram signals for the Problem move, as in examples (4) and (5), respectively.

However, it is interesting to note that while 狀況 ‘situation’ denoted the meaning of the situation or circumstance, it was preceded by the description of the problem, as illustrated in (4) below. In other words, in contrast to 時候 ‘time’ in (5), which was used to establish the context for the following discourse, 狀況 ‘situation’ in fact more often functioned as a shell noun (Schmid, 2000) or an evaluation carrier (Mahlberg, 2005) that marked a problematic situation.

(2) Example of 的問題 ‘problem of’

眼科醫師最擔心除了近視外，更擔憂高度近視、兩眼視差大的問題。

‘In addition to myopia, ophthalmologists are more concerned about the problem of high myopia and a significant difference in vision between eyes.’

(3) Example of 的原因 ‘reason of’

而造成這種可怕情況的原因，就是「快樂缺氧」，又稱為「隱形缺氧」。

‘What causes such a horrifying situation to happen is “happy hypoxia,” also known as “silent hypoxia”.’

(4) Example of 的狀況 ‘situation of’

雖然這些人可能還沒有出現器官實質上的損害，但其實已經出現功能上變差的狀況。

‘Although the organs of these people have not displayed physical damage, they actually have shown functional decline.’

(5) Example of 的時候 ‘time of’

不過有些人刷牙的時候都沒有流血，但就是牙齦腫腫的，壓到也會有點痛，搞不清楚自己現在到底是發生了什麼問題。

‘However, some people do not bleed when brushing their teeth. It’s just that their gums are swollen and would hurt when pressed. They aren’t really clear what has happened to them exactly.’

As for the medical terms that served as signals for the Problem component, many of them were related to physicians, such as 科醫師 ‘doctor from the division of’ and 主治醫(師) ‘attending physician’, as shown in (6) and (7), respectively. This may have been partly because the journalists often quoted the opinions of practicing physicians as trustworthy sources in the Problem move, as illustrated in the following extracts. These doctor-related trigram expressions thus appeared to be not only signals for the Problem element but also indicators of the

journalistic genre and the medical topic.

(6) Example of 科醫師 ‘doctor from the division of’

精神科醫師指出：「一般人看了天災人禍感到焦慮、不安、擔憂，也會出現『急性壓力障礙』」。

‘A psychiatrist points out that when witnessing disasters and accidents, people mostly would feel anxious, disturbed and worried and would develop acute stress disorder.’

(7) Example of 主治醫(師) ‘attending physician’

長安神經內科醫療中心主治醫師陳惠萱表示，很多人常常會有「慢性疲勞」的問題。

‘Huixuan Chen, attending physician at the Neurology Therapy Center of Everan Hospital, notes that many people would have the problem of chronic fatigue.’

Another group of medical terms that was often found in this move included nouns that indicated a problem, such as 症狀 ‘symptom’ and 的疾病 ‘disease of’, as shown in (8) and (9), respectively, below. These terms were often utilized to further explain the problem in question, as illustrated in the following examples. This group of frequent trigram expressions demonstrated that in addition to more general abstract nouns, such as problem and risk, nouns that were more topic specific yet less explicitly related to the concept of a problem also served similar functions. This highlighted the importance of examining and analyzing a rhetorical structure like the PS pattern in the context of a specific genre.

(8) Example of 症狀, ‘symptom,’

「腳痛」是很多人常常會有的症狀，可能是因為肌肉拉傷、抽筋，或是筋膜炎等原因導致。

“‘Foot pain’ is a common symptom that many people may experience, which may happen due to muscle strains, cramps, fasciitis, or other causes.’

(9) Example of 的疾病 ‘disease of’

失眠這個問題，李信謙解釋：「它是一種『慢性狀況、急性惡化』的疾病。」

‘Regarding the problem of insomnia, Xinqian Li explains that it is a chronic disease that may deteriorate acutely.’

In addition to nominal sequences, many of the verb-based trigram expressions were also more likely to be found in the Problem move. One type of verb that frequently occurred in this move was reporting verbs, such as 說 ‘say’ and 表示 ‘indicate’. Similar to and along with physician-related terms, reporting verbs were often used by the journalists of health news to cite an authoritative source in the Problem element, as exemplified in (10) below. One of the trigram signals contained both the noun 醫師 ‘physician’ and the reporting verb 說 ‘say’, as exemplified in (11) below. Again, this pattern showcased the characteristics of the health news genre, in addition to being a signal for the Problem move.

(10) Example of 表示, ‘indicate,’

中國醫藥大學新竹附設醫院婦產科醫師許馨予表示，許多人對於哺乳議題有迷思。
 ‘Xinyu Hsu, obstetrician at China Medical University Hsinchu Hospital, notes that many people have myths about breastfeeding.’

(11) Example of 醫師說 ‘doctor says’

蔡醫師說，多年前因對此疾病認知不足，泛視神經脊髓炎常被診斷成多發性硬化症。
 ‘Dr. Tsai says that many years ago, due to the lack of knowledge about this disease, neuromyelitis opticaspectrum disorder was often diagnosed as multiple sclerosis.’

Other types of verbs that tended to occur in the Problem move included modals such as 可能 ‘may’ and 會 ‘will’, causal verbs such as 造成 ‘cause’, and change-of-state verbs such as 出現 ‘appear’, as illustrated in (12), (13), and (14), respectively, below. In the data, all of these verbs were mostly utilized to introduce the problem into the discourse, as illustrated in the following examples. These frequent verbs and the trigram sequences in which they occurred displayed a strong negative semantic prosody (Sinclair, 1991), while the verbs themselves did not appear to be semantically negative (cf. Stubbs, 1995; Tao, 2003). Also noteworthy is the prevalence of the epistemic modal 可能 ‘may’, which helped express uncertainty in the medical discourse (cf. Lin, 2019) and, as the data showed, in the Problem move of health news.

(12) Example of 可能會 ‘possibly will’

但隨著年紀增長，或是不當使用膝蓋，膝關節可能會提早退化。

‘But as one ages or uses their knees incorrectly, their knee joints may possibly start to deteriorate at an earlier age.’

(13) Example of 會造成 ‘will cause’

再加上精緻飲食、高糖高油的飲食、高血壓高血脂等慢性疾病的影響，不只讓大腦變得遲鈍，甚至可能會造成「失智症」。

‘Along with the impact of refined foods, high-sugar and high-fat diets and chronic diseases such as hypertension and high cholesterol, it may not only cloud the brain but also lead to “dementia”.’

(14) Example of 會出現 ‘will appear’

皮膚科醫師提醒，使用 3C 產品除了要擔心用眼過度、近視外，也可能會出現皮膚症狀！

‘A dermatologist warns that in addition to overworked eyes and myopia, using electronic devices may also lead to the emergence of skin conditions!’

In agreement with the findings of previous studies, connectives that indicated the Reason-Result relationship, such as 所以 ‘so’ and 因為 ‘because’, as in (15) and (16), respectively, and adverbials of contrast, such as 其實 ‘actually’ and 甚至 ‘even’, were found to be signals for the Problem element (Flowerdew, 2008; Charles, 2011). The two adverbials exemplified in (17) and (18), respectively, allowed the writer to direct the reader’s attention to the “real problem” or the potential harm that may be caused by the problem in question. In contrast, less often reported in the literature and yet found in the frequent trigram signals for the Problem move was the adverbial 通常 ‘usually’. As illustrated in (19), the trigram sequence served as a hedge for the following clause. Similar to the modal verb 可能 ‘may’, which was discussed earlier, the high frequency of the adverbial 通常 ‘usually’ may have also reflected the uncertainty inherent in medical science and discourse (Lin, 2019).

(15) Example of 所以 ‘so’

而有些人在一開始只是「偶發頭痛」，常容易被忽略，所以會有 2% 左右的人演變成慢性頭痛。

‘And at the beginning, some people only have “occasional headaches” and thus overlook them, so approximately 2% of the people would develop chronic headache.’

(16) Example of 是因為 ‘be because’

《歐洲心臟病學雜誌》指出有四分之一的心臟病發作是因為高血壓所引起。

‘European Journal of Cardiology points out that a quarter of heart attacks were induced

due to hypertension.’

- (17) Example of adverbial 其實 ‘, actually’

林明秀也提醒，其實會產生疤痕，並不是傷口大小來決定的，而是傷口的深度及嚴重度。

‘Mingxiu Lin also emphasizes that the formation of scars in fact depends not on the size of the wound but on its depth and severity.’

- (18) Example of 甚至 ‘, even’

然而，中軸性脊椎關節炎與運動過度的狀況不同，更無法透過休息改善，甚至越久不動會越嚴重。

‘However, axial spondyloarthritis is different from overexercise. It cannot be improved by rest and may even be worsened by being sedentary.’

- (19) Example of 通常 ‘, usually’

另外，通常有狐臭的人，還容易伴隨有多汗症。

‘In addition, usually, people with armpit odor tend to have hyperhidrosis as well.’

Finally, the Problem move was also signaled by less common types of trigram expressions, such as the noun modifiers 常見的 ‘commonly seen’ and 這樣的 ‘such; of this kind’, and topic introducers, such as 像是 ‘such as’ and 來說 ‘as for; with respect to’. Despite the variety in meaning and parts of speech, these trigram signals served as discourse management functions, such as topicalizing, as in the examples of 常見的 ‘commonly seen’ in (20) and 來說 ‘as for’ in (21), introducing examples, as in the example of 像是 ‘such as’ in (22), and referring, as in the example of 這樣的 ‘such; of this kind’ in (23):

- (20) Example of 常見的 ‘commonly seen’

濕疹最常見的是搔癢與起疹子。

‘The most common (symptoms) of eczema are itches and rashes.’

(21) Example of 來說, 'as for,'

但對於耳鼻喉科的醫師來說，雖然喉嚨卡卡是一個小症狀，卻是很多疾病的共同警訊。

'But for ENT doctors, although a minor symptom, a sore throat is the common warning sign of several diseases.'

(22) Example of 像是, 'such as'

靜脈曲張常發生在久坐久站的職業類別，像是廚師、櫃姐、老師、保全等常受此困擾。

'Individuals with jobs that require them to sit or stand for a long time often develop varicose veins. Cooks, salesclerks, teachers, and security guards, for example, are often bothered by this condition.'

(23) Example of 這樣的, 'such; of this kind'

神經內科醫師解釋，這樣的情況容易發生在中年或已停經的女性。

'A neurologist explains that such conditions are often found in middle-aged or menopausal women.'

With respect to the functional distribution of trigrams in the Problem move, as can be seen in Table 5 below, most of the trigrams were used to serve the discourse organizing and referring functions, and only a couple of modal verb (e.g., 可能會 'probably will') and adverb (e.g., 其實 'actually') trigrams were used to serve stance functions. This may have been partly due to the written nature of the online health news articles (cf. Biber *et al.*, 2004), which is in correspondence with the findings of previous studies on lexical bundles in Chinese journalistic texts (Hsu, 2021). Another distinguishable feature that was observed in this list was that several trigrams in the discourse organizing category were employed to mark the information source, a genre-specific characteristic discussed earlier (cf. Biber & Gray, 2013). A number of trigrams were employed to introduce a topic for later discussion, such as 的原因 'reason of', 會造成 'will cause', and 來說, 'as for,'. Moreover, trigrams containing nouns that were semantically associated with the Problem move, such as 問題 'problem' and 風險 'risk', that identified the focus of the article or move mostly fell under the referring category. Trigrams such as 臨床上 'clinically' and 這樣的 'such; this type of' were used to specify the attribute of a topic, while trigrams such as 通常, 'usually' and 的時候 'time of' referred to a time-related concept (e.g., frequency) or point in the texts.

Table 5. Functional categories of the trigram signals for the Problem move

Types	Examples
I. STANCE	
A. Epistemic stance	可能會 ‘probably will’ ，可能 ‘, probably’
B. Attitudinal stance	，其實 ‘, actually’ ，甚至 ‘, even’
II. DISCOURSE ORGANIZING	
A. Information source	說：「 ‘say’ 醫師說 ‘doctor says’ 表示， ‘indicate,’ 科醫師 ‘doctor of the Division of’ 主治醫 ‘attending physician’ 治醫師 ‘attending physician’ 科主治 ‘attending physician at the Division of’
B. Topic management	的原因 ‘reason of’ 會造成 ‘will cause’ 會出現 ‘will appear’ ，像是 ‘, such as’ 來說， ‘as for,’
C. Discourse connection	，所以 ‘, so’ 是因為 ‘be because’
III. REFERRING	
A. Identification/focus	的問題 ‘problem of’ 問題。 ‘problem.’ 的風險 ‘risk of’ 問題， ‘problem,’ 的狀況， ‘situation of,’ 症狀， ‘symptom,’ 疾病， ‘disease,’ 的疾病 ‘disease of’
B. Specification of attributes	臨床上 ‘clinically’ 這樣的 ‘such; this type of’
C. Time reference	，通常 ‘, usually’ 的時候 ‘time of’

4.3.2 Signals for the Response Move

Despite the fact that more instances of the Response move were found in the corpus, fewer types of trigram signals were retrieved compared with those for the Problem move. The more frequent trigram signals for the Response move included those that involved abstract nouns (e.g., 情況 ‘situation’ and 方式 ‘method’), medical terms (e.g., 傷口 ‘wound’ and 治療 ‘cure; treatment’), modal verbs (e.g., 可以 ‘can’ and 有可能 ‘having the possibility; probably’), connectives (e.g., 如果 ‘if’, 因為 ‘because’, and 或是 ‘or’), and others (e.g., 越來越 ‘more and more’), as displayed in Table 6:

Table 6. Structural categories of the trigram signals for the Response move

Abstract Nouns	Medical Terms	Modal Verbs
的情況 ‘situation of’ 的方式 ‘method of’	的傷口 ‘wound of’ 的治療 ‘treatment of’	，可以 ‘, can’ 就可以 ‘just can’ 有可能 ‘having the possibility; probably’
Connectives	Others	
，如果 ‘, if’ ，因為 ‘, because’ ，或是 ‘, or’	越來越 ‘more and more’	

Unlike the Problem move, in which a Chinese counterpart for the abstract noun problem was found to be one of the prevalent signals, the Response move was not signaled by trigram expressions that involved nouns that denoted a solution in Chinese. This was partly due to the fact that the nominal counterpart of 解法 or 解方 ‘solution’, for example, is not as commonly used in Chinese. Instead, nouns that were less explicitly linked to the function of the move, such as 方式 ‘method; approach’, were involved in the sequences, as in (24). On the other hand, circumstantial abstract nouns, such as 情況 ‘situation’, were also found to be frequent in this move. However, as illustrated in (25) below, despite being a near-synonym of 狀況 ‘situation’ discussed earlier, 情況 ‘situation’ was not used to present a problem; instead, it was deployed to mark the context in which the recommended solution should be implemented or would be appropriate.

(24) Example of 的方式 ‘method of’

如果傷口略大，可以透過輕壓的方式幫助止血。

‘If the wound is fairly big, you can try (the method of) pressing lightly on the wound to stop bleeding.’

(25) Example of 的情況 ‘situation of’

而若遇到疫情較嚴重的情況下，醫師仍建議患者應規律就醫回診、用藥。

‘Even if the situation of the pandemic becomes more serious, physicians still recommend that patients should visit the doctor and take medication regularly.’

With respect to medical terms, in stark contrast to the Problem move, no human-related nouns were found to serve as signals for the Response move. This may have been in part because the Response move was typically subsequent to the Problem move, in which the main source of information was mentioned the first time and thus with the full title listed, and as a result, the source was usually referred to solely by name in the Response move, such as the more frequent health- and medicine-related ‘cure; treatment’ in (26), and nouns that referred to things that could or should be treated in particular ways, such as 傷口 ‘wound’ as exemplified in (27) below:

(26) Example of 的治療 ‘treatment of’

除了靠藥物的治療外，建議也需規律運動。

‘In addition to medical treatment, regular exercise is also recommended.’

(27) Example of 的傷口 ‘wound of’

這種類型傷口可能需要縫合，尤其是傷及真皮層的傷口更要小心護理，應該就醫評估，是否要進一步治療。

‘This type of wound may require sutures. Wounds that damage the dermis in particular should be treated with extra care. One should consult a doctor to assess if further treatment is needed.’

Similar to the Problem move, the Response move was marked by trigram sequences that involved modal verbs. However, rather than being signaled by epistemic modals such as 可能 ‘may’ and 會 ‘will’, the Response move more often featured the modal verb 可以 ‘can’, as illustrated in (28) and (29) below. As pointed out by researchers such as Wang (2018), in addition to its “ability” reading, 可以 ‘can’ can also be used to present a suggestion or recommendation, as illustrated in (28) and (29), which explains why this modal verb was a frequent signal for the Response move whose communicative function was to recommend a way to solve the problem in question.

- (28) Example of 可以 ‘can’

所以，如果想選用抗沾黏產品，可以注意幾個要點。

‘Therefore, you can pay attention to a few key points when choosing anti-adhesion products.’

- (29) Example of 就可以 ‘then can’

這時候就可以透過現在很流行的筋膜槍，來讓我們身上的肌肉放鬆。

‘Then we can use the now very popular fascia gun to relax our muscles.’

It should be noted that although 可能 ‘possibility’ was also found in one of the trigram signals for the Response move, it was different from the use of 可能 ‘may’ in the Problem move. First, it was used as a noun-like item in the Response move because it was preceded by the verb 有 ‘to have’. Second, in lieu of identifying the key component of the move (i.e., the solution), the trigram sequence 有可能 ‘having the possibility; probably’ presented the contextual information that supported the construction of the Response move, as exemplified in (30) below:

- (30) Example of 有可能 ‘having the possibility; probably’

孩子大一點，半夜醒來，有可能是因為做夢或白天刺激過多造成，並不一定是沒吃飽。

‘As the child grows older, if they wake up at midnight, it may be probably due to dreaming or overstimulation during the daytime, not necessarily due to not having enough food.’

As for connectives that signaled the Response move, the markers of reason, such as 因為 ‘because’, were found to have achieved this function. However, more interestingly, connectives that were less often reported in the literature, such as the conditional marker 如果 ‘if’ and the disjunctive marker 或是 ‘or’, were frequent in the Response element in the corpus. The high frequency of these two connectives may have been due to the fact that the Response part of a medical news article involved the act of giving advice or making suggestions. As shown in (31) and (32) below, these two markers, respectively, often introduced the solution or response to the problem in question. This pattern, albeit distinct from Flowerdew’s (2008) findings, corresponded with the observations in previous research, that conditionals and expressions indicating alternativeness were often employed to support the advice-giving act in the discourse (Hsieh, 2019; Lin, 2019).

(31) Example of 如果 ‘if’

但在夏天或是劇烈運動後，如果臉部感到很油膩則可以再多洗一次。

‘But in summer or after strenuous exercise, if your face still feels greasy, then you can wash it one more time.’

(32) Example of 或是 ‘or’

民眾日常飲食建議多攝取富含維生素 B 群、C、D 的蔬果，如：蘋果、酪梨、奇異果等，或是含有豐富礦物質的深綠色蔬菜。

‘People are recommended to consume more vegetables and fruits that are rich in Vitamins B, C, and D, such as apples, avocados, and kiwis, or leafy greens, which contain abundant minerals.’

Finally, the comparative modifier 越來越 ‘more and more’, similar to the abstract noun 情況 ‘situation’ and the conditional marker 如果 ‘if’ discussed earlier, also functioned to present the condition or context for the solution presented, as illustrated in (33) below:

(33) Example of 越來越 ‘more and more’

如果針眼越來越大顆，還是直接去看眼科醫生最為有用。

‘If the sty gets bigger and bigger, it is still better to just go to an ophthalmologist.’

Compared with the Problem move, the Response move only showed a slight preference for the referring function, with much less diverse discourse organizing trigrams, as illustrated in Table 7 below. Stance functions in this move were served largely by epistemic modal verb trigrams, such as 就可以 ‘just can’ and 有可能 ‘having the possibility; probably’, whereas discourse organizing functions were achieved by trigrams containing grammatical connectives, such as 如果 ‘if’, 因為 ‘because’, and 或是 ‘or’. Similar to the pattern discussed in the previous section, noun-based trigrams that more explicitly indicated the Response move, such as 的方式 ‘method of’, 的傷口 ‘wound of’, and 的治療 ‘treatment of’, tended to be used to serve the referential function, suggesting the importance of referring trigrams in signaling the moves in the PS discourse structure in this genre. Finally, as mentioned earlier, while 狀況 ‘situation’ and 情況 ‘situation’ appeared to be near-synonymous, they in fact played distinctive roles in this genre of text. In the Problem move, trigrams that contained 狀況 ‘situation’ functioned to identify the focus of the move (i.e., the problem). In contrast, trigrams that involved 情況 ‘situation’ only referred to contextual information, such as the

time and condition, in the Response move. This showed the importance of a more qualitative, function-oriented analysis of the results.

Table 7. Functional categories of the trigram signals for the Response move

Types	Examples
I. STANCE	
A. Epistemic stance	, 可以 ‘, can’ 就可以 ‘just can’ 有可能 ‘having the possibility; probably’
II. DISCOURSE ORGANIZING	
A. Discourse connection	, 如果 ‘, if’ , 因為 ‘, because’ , 或是 ‘, or’
III. REFERRING	
A. Identification/focus	的方式 ‘method of’ 的傷口 ‘wound of’ 的治療 ‘treatment of’
B. Specification of attributes	越來越 ‘more and more’
C. Time reference	的情況 ‘situation of’

5. Conclusion

As one of the first projects that examined the Problem-Solution pattern in Chinese discourse, the current study constructed a small corpus of online health news articles annotated with elements of the PS textual pattern. An n-gram approach was then adopted in an attempt to search for the linguistic signals in this genre for each move in the PS pattern, and the findings were fruitful. The results showed that high-frequency trigram expressions retrieved from the corpus displayed some of the same characteristics reported in prior studies, such as indicating the Reason-Result relationship, marking contrast, and signaling the key components of the PS pattern, including Problem and Solution. On the other hand, features that have been mentioned in the literature, such as quotative devices and medically relevant terms, were also frequently used in the recurrent trigram expressions, and these features helped achieve functions that were highly relevant to the genre of online health news.

A closer look at the Problem and Response components in the online health news articles revealed that each of the key moves in the PS pattern preferred particular trigram sequences. For example, the Problem move in the dataset was often signaled by abstract nouns explicitly

or implicitly linked to the notion of problems and by devices for citing an authoritative source, such as a reporting verb and the noun 主治醫(師) ‘attending physician’ in Chinese. This move also tended to be marked by sequences that contained modals that conveyed prediction and uncertainty or verbs that indicated a (usually negative) change of state with negative semantic prosody. Regarding the functions that the trigrams served, the signals in the Problem move tended to fall under the discourse organizing and referring categories. A number of the expressions were used to cite the information source or refer to move-related topics, such as problems, risks, and diseases.

On the other hand, trigram signals for the Response move often involved linguistic features that enabled the writer to present suggestions or advice, such as dynamic verbs, conditionals, and disjunctive markers. Moreover, as the counterpart of the noun “solution” in Chinese is not frequently used, the abstract nouns found in the trigram signals were less explicitly related to “solution,” which displayed a language-specific characteristic. In contrast to the Problem move, the Response move displayed only a mild preference for referring trigrams. The stance and discourse organizing trigrams in the Response move also served less diverse functions. However, similar to the Problem move, the referring trigrams in the Response move played an important role in signaling the rhetorical structure.

Given the findings summarized above, the present study has a number of implications. First, while most of the previous studies on the PS pattern adopted the methods of keyword analysis to investigate the linguistic cues of the move structure, the current research demonstrated that the n-gram or multiword approach is a potential alternative framework for identifying lexicogrammatical signals for particular rhetorical functions and components, as this approach enabled us to examine linguistic resources beyond the constraints of traditional, prescriptive definitions of linguistic units (Cortes, 2013; Tian *et al.*, 2020).

Moreover, the findings presented in this article also exemplified the interaction between a rhetorical structure, such as the PS pattern, and a journalistic genre, in the case of this study, online health news articles written in Mandarin Chinese. As observed in the trigram signals retrieved from the corpus, while a popular rhetorical structure, especially in writing, the PS pattern featured linguistic markers that were characteristic of the genre for which the pattern was specifically utilized. The results also suggested that a few of the move or component signals were formulated to achieve communicative functions that were prominent or specific to the genre, such as citing a source or giving advice. These communicative acts in turn helped the writer to achieve the rhetorical function(s) of each move, such as identifying the problem and presenting the solution (cf. Belmonte, 2009; Ali, 2013). This points to the importance of putting the PS pattern in the context of a particular genre and of including more types of text in research on the PS structure.

On the basis of the current research, a number of directions for future research were

identified. First, a larger number of texts of the same (or different) genre(s) from different sources should be collected to see whether and to what extent the conclusions drawn in this article still apply and whether trigram expressions are the best unit of analysis. Second, in addition to the analysis of the relationship between frequent trigrams and the components of the Problem-Solution pattern, future studies should look into the distribution of communicative acts in each move of this genre and the mapping between communicative acts and n-gram sequences, as a few previous studies have examined (Belmonte, 2009; Ali, 2013). Lastly and probably most excitingly, the discourse data annotated for and the concluding findings in the current research project will serve as the foundation for training models for the automatic detection and annotation of the PS pattern and for developing other related applications, such as chatbots for medical purposes and algorithms that translate or even produce journalistic texts on health-related topics.

References

- Ali, A. M. (2013). Combining problem-solution categories and communicative acts: An analysis of Malaysian and British business journalistic texts. *World Applied Sciences Journal*, 21, 174-185. <https://doi.org/10.5829/idosi.wasj.2013.21.stl.2152>
- Belmonte, I. A. (2009). Toward a genre-based characterization of the problem–solution textual pattern in English newspaper editorials and op-eds. *Text & Talk*, 29(4), 393-414. <https://doi.org/10.1515/TEXT.2009.021>
- Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371-405. <https://doi.org/10.1093/applin/25.3.371>
- Biber, D., & Gray, B. (2013). *Discourse characteristics of writing and speaking task types on the TOEFL IBT® test: A lexico-grammatical analysis* (RR-13-04, TOEFLiBT-19). ETS Research Report Series.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135-146. https://doi.org/10.1162/tacl_a_00051
- Cavnar, W. B., & Trenkle, J. M.. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, Vol. 161175.
- Charles, M. (2011). Adverbials of result: Phraseology and functions in the problem–solution pattern. *Journal of English for Academic Purposes*, 10(1), 47-60. <https://doi.org/10.1016/j.jeap.2011.01.002>
- Cortes, V. (2013). The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes*, 12(1), 33-43. <https://doi.org/10.1016/j.jeap.2012.11.002>

- De Coninck, D., d'Haenens, L., & Matthijs, K. (2020). Forgotten key players in public health: News media as agents of information and persuasion during the COVID-19 pandemic. *Public Health*, 183,65-66. <https://doi.org/10.1016/j.puhe.2020.05.011>
- Escalante, H. J., Solorio, T., & Montes-y-Gómez, M. (2011). Local histograms of character n-grams for authorship attribution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 288-298.
- Flowerdew, L. (2003). A combined corpus and systemic-functional analysis of the problem-solution pattern in a student and professional corpus of technical writing. *TESOL Quarterly*, 37(3), 489-511. <https://doi.org/10.2307/3588401>
- Flowerdew, L. (2008). *Corpus-based analyses of the problem-solution pattern: A phraseological approach*. John Benjamins.
- Fox, S., & Duggan, M. (2013). *Health online 2013*. PEW RESEARCH CENTER
- Fu, J., Liu, P., Zhang, Q., & Huang, X.-J. (2020). Is Chinese word segmentation a solved task? Rethinking neural Chinese word segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5676-5686. <https://doi.org/10.18653/v1/2020.emnlp-main.457>
- Ghenai, A., & Mejova, Y. (2018). Fake cures: User-centric modeling of health misinformation in social media. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 1-20. <https://doi.org/10.1145/3274327>
- Giannakopoulos, G., & Karkaletsis, V. (2009). N-gram graphs: Representing documents and document sets in summary system evaluation. In *Proceedings of Text Analysis Conference TAC 2009*.
- Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4), 403-437. <https://doi.org/10.1075/ijcl.13.4.02gri>
- Handford, M., & Matous, P. (2015). Problem-solving discourse on an international construction site: Patterns and practices. *English for Specific Purposes*, 38, 85-98. <https://doi.org/10.1016/j.esp.2014.12.002>
- Heffernan, K., & Teufel, S. (2018). Identifying problems and solutions in scientific text. *Scientometrics*, 116(2), 1367-1382. <https://doi.org/10.1007/s11192-018-2718-6>
- Hoey, M. (1983). *On the surface of discourse*. Allen & Unwin.
- Hoey, M. (2001). *Textual interaction: An introduction to written discourse analysis*. Psychology Press.
- Hsieh, C.-L. (2005). Modal verbs and modal adverbs in Chinese: An investigation into the semantic source. *UST Working Papers in Linguistics*, 1, 31-58.
- Hsieh, C.-Y. C. (2019). *Language, intersubjectivity, and institutional interaction: Advice-giving directives in Taiwan EFL writing tutorials* (Doctoral dissertation). National Taiwan University.
- Hsieh, C.-Y. C. (2020). Meaning in repair: The abstract noun yisi 'meaning/intention' in the management of intersubjectivity in Mandarin conversation. *Taiwan Journal of Linguistics*, 18(2), 39-88. [https://doi.org/10.6519/TJL.202007_18\(2\).0002](https://doi.org/10.6519/TJL.202007_18(2).0002)

- Hsu, C.-C. (2021). The structure and function of lexical bundles in Chinese conversation and news. *Taiwan Journal of Chinese as a Second Language*, 22, 69-96. [https://doi.org/10.29748/TJCSL.202106_\(22\).0003](https://doi.org/10.29748/TJCSL.202106_(22).0003)
- Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), 4-21. <https://doi.org/10.1016/j.esp.2007.06.001>
- Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development*, 48(4), 63-85. <https://doi.org/10.1007/BF02300500>
- Kachru, B. B. (1992). *The other tongue: English across cultures*. University of Illinois Press.
- Kanaris, I., Kanaris, K., Houvardas, I., & Stamatatos, E. (2007). Words versus character n-grams for anti-spam filtering. *International Journal on Artificial Intelligence Tools*, 16(06), 1047-1067. <https://doi.org/10.1142/S0218213007003692>
- Kumari, S., Reddy, H. K., Kulkarni, C. S., & Gowthami, V. (2021). Debunking health fake news with domain specific pre-trained model. *Global Transitions Proceedings*, 2(2), 267-272. <https://doi.org/10.1016/j.gltp.2021.08.038>
- Kuta, M., & Kitowski, J. (2014). Optimisation of character n-gram profiles method for intrinsic plagiarism detection. In *Proceedings of the 13th International Conference on Artificial Intelligence and Soft Computing (ICAISC 2014)*, 500-511. https://doi.org/10.1007/978-3-319-07176-3_44
- Li, L., Franken, M., & Wu, S. (2020). Bundle-driven move analysis: Sentence initial lexical bundles in PhD abstracts. *English for Specific Purposes*, 60, 85-97. <https://doi.org/10.1016/j.esp.2020.04.006>
- Lin, W.-H. (2019). Expressing uncertainty with conditionals in medical discourse: A comparison across genres. In: Tao, H., Chen, HJ. (eds) *Chinese for Specific and Professional Purposes. Chinese Language Learning Sciences*. Springer. https://doi.org/10.1007/978-981-13-9505-5_10 213-243.
- Liu, M.-C., & Chiang, T.-Y. (2008). The construction of Mandarin VerbNet: A frame-based study of statement verbs. *Language and Linguistics*, 9(2), 239-270.
- Mahlberg, M. (2005). *English general nouns: A corpus theoretical approach*. John Benjamins.
- Ng, H. T., & Low, J. K. (2004). Chinese part-of-speech tagging: One-at-a-time or all-at-once? Word-based or character-based? In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 277-284.
- Ratanakul, S. (2017). A study of problem-solution discourse: Examining TED talks through the lens of move analysis. *LEARN Journal: Language Education Acquisition Research Network*, 10(2), 25-46.
- Ratanakul, S. (2018). A move analysis of problem-solution discourse: A pedagogical guide for opinion and academic writing. *Arab World English Journal*, 9(3), 233-247. <https://doi.org/10.24093/awej/vol9no3.16>
- Schmid, H.-J. (2000). *English abstract nouns as conceptual shells: From corpus to cognition*. De Gruyter Mouton.

- Scott, M. (2000). Mapping key words to problem and solution. In M. Scott & G. Thompson (Eds.), *Patterns of Text: In honour of Michael Hoey* (pp. 109-127). John Benjamins.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
- Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and eap writing development: Lexical bundles in the TOFEL iBT writing section. *Journal of English for Academic Purposes*, 12(3), 214-225. <https://doi.org/10.1016/j.jeap.2013.05.002>
- Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, 2(1), 23-55. <https://doi.org/10.1075/foL.2.1.03stu>
- Tao, H. (2003). Toward an emergent view of lexical semantics. *Language and Linguistics*, 4(4), 837-856.
- Tian, Y., Song, Y., Ao, X., Xia, F., Quan, X., Zhang, T., & Wang, Y. (2020). Joint Chinese word segmentation and part-of-speech tagging via two-way attentions of auto-analyzed knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8286-8296. <https://doi.org/10.18653/v1/2020.acl-main.735>
- Tsirintani, M. (2021). Fake news and disinformation in health care—challenges and technology tools. In *Public health and informatics (Proceedings of MIE 2021)*, 318-321. IOS Press.
- Wang, P. Y. (2018). *A cognitive-pragmatic study on modal verbs of possibility in Chinese* (Doctoral dissertation). The Pennsylvania State University.
- Waszak, P. M., Kasprzycka-Waszak, W., & Kubanek, A. (2018). The spread of medical fake news in social media—The pilot quantitative study. *Health Policy and Technology*, 7(2), 115-118. <https://doi.org/10.1016/j.hlpt.2018.03.002>
- Waugh, L. R. (1995). Reported speech in journalistic discourse: The relation of function and text. *Text & Talk*, 15(1), 129-73. <https://doi.org/10.1515/text.1.1995.15.1.129>
- Wieting, J., Bansal, M., Gimpel, K., & Livescu, K. (2016). Charagram: Embedding words and sentences via character n-grams. In arXiv preprint arXiv:1607.02789
- Yue, M. (2006). Discursive usage of six Chinese punctuation marks. In *Proceedings of the COLING/ACL 2006 Student Research Workshop*, 43-48.
- 李松濤、許文怡 (2020)。科學傳播歷程中程序性知識特徵的框架探究-以飲食保健類科學研究新聞為例。 *科學教育學刊* , 28(2) , 143 - 168 。 [https://doi.org/10.6173/CJSE.202006_28\(2\).0003](https://doi.org/10.6173/CJSE.202006_28(2).0003) 。 [Lee, S. T. & Hsu, W.-Y. (2020). A Framing Exploration of Procedural Knowledge Characteristics in Science Communication Process: Examples of Research Based Science News Regarding Healthy Diet. *Chinese Journal of Science Education*, 28(2), 143-168.]
- 李松濤、鄔啟柔 (2017)。科學新聞傳播內容與模式之探究以飲食、疾病與醫藥類新聞為例。 *科學傳播論文集* 8 , 17 - 30 。 <https://doi.org/10.6930/scicomm.201705.0002> 。 [Lee, S. T. & Wu, C. R. (2017). An exploration on the content and model of science news communication: A case study of news stories related to food, disease, and medicine. *Collected Papers on Science Communication* 8, 17-30.]

葉蓉慧、黃暉超 (2017)。社群網路時代民眾對新聞回應:以臉書民眾對台灣流感議題回應為例。科學傳播論文集 8, 103 - 113。 <https://doi.org/10.6930/scicomm.201705.0006>。
[Yeh, J. H. B. & Huang, W. C. (2017). Readers' responses to news in the social media era: A case study of Facebook users' responses to flu-related issues in Taiwan. *Collected Papers on Science Communication* 8, 103-113.]

Appendix A. Statistics of the 58 trigrams with a Min-Max scaling value larger than 0.3

Trigrams	Frequency	Relative Frequency	Min-Max Scaling
(，如果)	113	0.000895724	1.0000
(說：「)	85	0.000673774	0.7500
(，所以)	80	0.000634141	0.7054
(，因為)	75	0.000594507	0.6607
(科醫師)	74	0.00058658	0.6518
(可以)	74	0.00058658	0.6518
(症狀，)	71	0.0005628	0.6250
(的時候)	70	0.000554873	0.6161
(攝護腺)	69	0.000546946	0.6071
(的問題)	69	0.000546946	0.6071
(，其實)	68	0.000539019	0.5982
(，或是)	68	0.000539019	0.5982
(表示，)	67	0.000531093	0.5893
(糖尿病)	64	0.000507312	0.5625
(問題，)	62	0.000491459	0.5446
(，甚至)	56	0.000443898	0.4911
(治療，)	55	0.000435972	0.4821
(可能會)	54	0.000428045	0.4732
(高血壓)	53	0.000420118	0.4643
(的患者)	53	0.000420118	0.4643
(主治醫)	52	0.000412191	0.4554
(治醫師)	52	0.000412191	0.4554
(就可以)	52	0.000412191	0.4554
(這樣的)	50	0.000396338	0.4375
(的症狀)	49	0.000388411	0.4286

(護腺癌)	48	0.000380484	0.4196
(是因為)	48	0.000380484	0.4196
(，可能)	47	0.000372558	0.4107
(有可能)	47	0.000372558	0.4107
(的原因)	46	0.000364631	0.4018
(，因此)	46	0.000364631	0.4018
(疾病，)	44	0.000348777	0.3839
(膝關節)	44	0.000348777	0.3839
(的狀況)	43	0.000340851	0.3750
(性皮膚)	43	0.000340851	0.3750
(的情況)	43	0.000340851	0.3750
(，通常)	43	0.000340851	0.3750
(皮膚炎)	42	0.000332924	0.3661
(的傷口)	41	0.000324997	0.3571
(疤產品)	40	0.00031707	0.3482
(醫師說)	40	0.00031707	0.3482
(問題。)	40	0.00031707	0.3482
(的治療)	40	0.00031707	0.3482
(會造成)	40	0.00031707	0.3482
(疤痕，)	40	0.00031707	0.3482
(釋：「)	39	0.000309144	0.3393
(解釋：)	39	0.000309144	0.3393
(越來越)	39	0.000309144	0.3393
(常見的)	37	0.00029329	0.3214
(的疾病)	37	0.00029329	0.3214
(的方式)	37	0.00029329	0.3214
(會出現)	36	0.000285363	0.3125
(疤痕的)	36	0.000285363	0.3125
(科主治)	36	0.000285363	0.3125

(臨床上)	36	0.000285363	0.3125
(，像是)	36	0.000285363	0.3125
(來說，)	35	0.000277436	0.3036
(的風險)	35	0.000277436	0.3036

Let Me Finish!—Speech Patterns of Interruptions in Chinese: A Corpus-based Study on Parliamentary Interpellations on Taiwan

Christian Schmidt* and Chia-Rung Lu⁺

Abstract

This corpus-based study investigated verbal interruptions during parliamentary interpellations based on official and publicly accessible transcriptions provided by the Legislative Yuan of the Republic of China (Taiwan). While interruptions have previously been understood as organizing turn-taking, as well as cues and speech markers, the results of this study suggest that interruptions have a dual nature. Interruption is incentivised by confrontational discourse strategies and realized by linguistic expressions, some of which are statistically significant and can be called keywords. Using open-source data to explore the linguistic features in the speech patterns of interruptions in institutional discourse, we first identified the word classes and keywords with significant frequency shifts between interrupted, interrupting, and regular sentences. Then, we associated the meanings of the keywords with offensive and defensive discourse strategies. The findings of this study indicate that interrupted sentences were more reflective of defensive discourse strategies, while interrupting sentences were associated with offensive ones. Moreover, conjunctions, adverbs, and pronouns played a more important role in the speech patterns of interruptions compared with their respective footprint in the lexicon. Conversely, nouns and verbs, with some exceptions, as well as adjectives, played a lesser role. We argue that the confrontational incentive structure in institutional debates creates certain linguistic patterns, mostly statistically significant frequency shifts of keywords in interrupted and interrupting sentences, and that these patterns might be useful in explaining interruption.

* Graduate Institute of Linguistics, National Taiwan University, Taipei, Taiwan.
E-mail: d05142004@ntu.edu.tw (orcid.org/0000-0002-6746-8762).

⁺ Graduate Institute of Linguistics, National Taiwan University, Taipei, Taiwan.
E-mail: chiarung@ntu.edu.tw (orcid.org/0000-0002-9275-7371).

The author for correspondence is Chia-Rung Lu.

Keywords: Speech Patterns, Spoken Chinese, Interruptions, Institutional Discourse

1. Introduction

Parliamentary discussions in Taiwan are reflective of the democratic nature of legislature. Audio-video recordings have well documented highly engaged lawmakers from all parties attacking opponents verbally and sometimes physically. Less studied, however, are the linguistic aspects of verbal confrontations during the institutional discourse. This study focused on verbal interruptions and their role in the nature of language. Due to the argumentative nature of parliamentary discussions, this type of discourse produced aggressive and extraordinary linguistic data, which revealed some specific properties of language, such as debate strategies, possible clausal boundaries, and context-tinged vocabulary.

Taking a corpus-based approach with a heightened focus on interruptions, this study aimed at addressing two main questions. First, which linguistic level explains interruptions better, parts of speech or keywords? In terms of parts of speech, we applied the automated tagging system developed by the Chinese Knowledge and Information Processing (CKIP) Lab. Keywords refer to the words that appeared significantly more often in interrupted and interrupting sentences compared with those in regular sentences. Second, how are keywords linked to discourse functions, and is there an underlying semantic relationship between the keywords and the discourse functions?

Finally, verbal interruptions do not happen in a vacuum. In agreement with Stainton (1987), certain situational contexts create an incentive structure for a confrontational style of conversation. Therefore, the notion of an “incentive structure” at least partly accounts for the explanation of interruptions.

2. Literature Review

2.1 Interruption

Previous research has shown that interruptions are caused by certain keywords, also called cues (Duncan, 1972; Wiemann & Knapp, 1975). Others have suggested that it is rather the intention and motivation of the interrupter that plays a key role (Oreström, 1983; Bazzanella *et al.*, 1991; Waltereit, 2002). Conversational analysts have considered interruptions a subtype of turn-taking, often with the implicit assumption that interruptions do not happen by chance but are linguistically marked (Sacks *et al.*, 1974).

Starting from this broader perspective, Ferguson (1977) presented a classification of turn-taking in conversations that distinguished between overlaps (i.e., the current speaker, despite the intervention, is determined to reach turn-completion) and a single interruption (i.e., the most common type), and that recognized interruptions as a mechanism of turn-taking along the lines

of illocutionary effects on the speaker and the hearer. Ferguson (1977) also mentioned other forms, such as smooth speaker switching (i.e., a change between two speakers with no interruptions) and silent interruptions (i.e., simple, silent interruptions that indicate that the current speaker should give up his or her turn). These other forms of interruptions require physical clues for identification, which was beyond the scope of this study.

As an early conversational analyst, Duncan (1972) also identified cues, such as turn-yielding cues, back-channel cues, and turn-maintaining cues, and construed them as triggers that indicated when turn-taking should take place. Wiemann and Knapp (1975) later expanded this list by adding turn-requesting cues. Oreström (1983) also claimed that interruptions cannot be satisfactorily described only with the help of formal criteria because a subjective element is always involved, and that there is no specific and unambiguous marker, grammatical or lexical, of turn-finality. Oreström (1983) suggested considering more factors to better classify interruptions and proposed a categorization based on grammatical boundaries and turn-taking, including loudness, speed, length, discourse content, floor winning, age and sex of the speaker, the manner of recording, and the ongoing speaker's reactions. Oreström (1983) further established a typology of reasons why interruptions happen and the interruption types observed in conversational practice, such as anti-communication (i.e., imperatives such as 'Stop that!'), protests (e.g., 'That's not true!'), and check-up questions (e.g., 'Why did you just say that?'). As Waltereit (2002) pointed out, this list was extended by Bazzanella (1991), who included the psychological element (i.e., interruptions that show emotional effect) and force majeure (i.e., interruptions that reflect that two speakers belong to different social power structures).

Signes (2000) continued along this line, that the kind of turn-taking resulting from an intervention also reveals the emotional orientation of the speakers toward the institutional character of the interaction, the internal social status and mind-set of the speakers, and the identities taking part in it. At odds with Hutchby (1996), Signes (2000) argued that interruptions are dualistic in nature, that is, cooperative and/or disruptive; that what counts as cooperative or disruptive is subjective, depending on what the speaker/listener thinks; and, finally, that the interpretation of an interruption is dyadic and intersubjective in nature, meaning that the interpretation is influenced not only by the participants of the conversation but also by the basic setting or type of conversation. In terms of classification, Signes (2000) categorized interruptions by function: interruptions, overlaps, and parenthetical remarks. Levinson (1983) added inadvertent overlaps and violative interruptions to that list.

More recently, Waltereit (2002), based on the earlier work of Jefferson (1978) and others, discussed interruptions in terms of a conversational practice, summarizing that interruptions are a normal part of the conversational practice and, to a certain extent, are tolerated if a speaker points to something extremely urgent or considers the current conversation irrelevant. Waltereit (2002) mentioned research by Tannen (1984) and Bazzanella (1991), who posited that

interruptions can even be regarded as a form of positive politeness if they are aimed at cooperative joint formulation.

In the search for causes of interruptions, links have been made to discourse markers. Nor (2012) demonstrated that discourse markers (Fraser, 1990) such as ‘well’, ‘now’, and ‘and’ are used as turn-initial interruptive devices and used Schegloff’s (2002) framework of what constitutes an interruption in turn-taking, with a focus on the functions of discourse markers. However, verbal interruptions in the Chinese-speaking context are still underresearched, especially interruptions related to institutional discourse.

2.2 Incentive Structure

As shown in the previous section, the literature on interruptions has identified certain formal interruption types linked to motivation, such as imperatives, protests, and check-up questions, but it is overly simplistic to argue that interruptions are either directly encouraged or caused by cues or discourse markers. Instead, Stainton (1987) provided some arguments in considering the incentive structure of interruptions, asserting that the distribution of interruption types is influenced by situational context, that the degree of social distance between the participants is an important factor, and that different degrees of social distance influence the frequency of interruptions. Stainton’s (1987) argument is important because it can be extended to include politically and socially constructed distance, such as that in different political parties and in pursuing different political goals.

In this study, we investigated how interruptions emerged in the context of political interpellations. We hypothesized that, specifically in the context of political interpellations, the underlying incentive structure promotes discourse strategies that include interruptions and other aggressive speech acts in order to create specific illocutionary effects and dominance over the discourse opponent and to undermine the opponent’s credibility. The literature on interruptions in political discourse is rather limited, but Goldberg (1990) held that although not synonymous with power, some interruptions may indeed signal power, rapport, and cooperation, differentiating in general between power interruptions and non-power interruptions. According to Goldberg (1990), power interruptions can be understood as a power play between two interlocutors, in which the interrupter breaks in and cuts off the speaker as a way to display social power. Such a display of social power is understood as an act of competition, or even conflict, and is regarded as impolite, rude, and forthrightly hostile or disrespectful toward the speaker and his or her message.

This line of reasoning has drawn attention to one important element in the incentive structure: power. Power is a social construct and a quantifiable factor, at least nominally, in political interpellations because speakers belong to different parties and different groups within the political system, either in the government or as legislators—most typically, members of the

parliament are conducting an interpellation, and a member of the government is answering. This setting makes clear Goldberg's (1990) differentiation between power and non-power interruptions. In interpellations, the incentive structure rewards speech actions that aim at questioning, showing power, ridiculing, and pressing. Furthermore, Hutchby (1996) distinguished between cooperative and non-cooperative interruptions with the idea that an interruption can be purposeful and can be used as a rhetorical device, instead of being just passively triggered by cues. The current study agrees with the notion of "interruption on purpose"; therefore, interrupted and interrupting sentences were examined separately.

In summary, an incentive structure shapes the motivation of participants to communicate in a certain way and hence is more directly associated with discourse strategies, and even linguistic expressions, than with context or intersubjectivity. We hypothesized that interruption effects could be observed at, albeit not fully explained by, the parts-of-speech level.

2.3 Defining Interruption Incentive Structure

In this study, we considered sentences "interrupted" if marked as such in the official transcripts provided by the official website of the Legislative Yuan of the Republic of China (Taiwan). Specifically, interrupted sentences were those that were explicitly marked as incomplete, using a set of three dots to indicate an ellipsis (...) at the end of an utterance.

We regarded any sentence that directly followed an interrupted utterance an "interrupting" sentence; hence, each interrupted utterance had an interrupting counterpart. In a few cases, an interrupting sentence was interrupted by a following sentence, and those special cases were listed as both interrupted and interrupting. Due to their small number, the impact on the analysis was negligible.

Any sentence that was neither interrupted nor interrupting was defined as a "regular" sentence that ended with a full stop (.), an exclamation mark (!), or a question mark (?), rather than an interrupted/interrupting sentence that ended with an ellipsis (...). Moreover, since our definition of interrupted was limited to sentences, grammatically defined as a complete syntactic unit in written or spoken form, and did not include utterances, which generally refer to any number of words spoken (uttered) during a conversation, utterances were irrelevant in our analysis. However, because the data was extracted from the official transcripts of the Legislative Yuan of Taiwan, we therefore used the terms "sentence" and "utterance" interchangeably in the discussion presented in this paper.

3. Methodology

3.1 Corpus

The corpus of this study was built from the transcribed recordings of various official meetings of Taiwanese lawmakers and ministers, which are publicly available on the official website of the Legislative Yuan of the Republic of China, Taiwan (立法院中部辦公室).¹ The transcribed materials included in the corpus were extracted from the documents and records shown in Tables 1 and 2:

Table 1. Minutes of the 5th meeting of the 2nd session of the 9th Legislative Yuan

No.	Dates	Meetings or Topics
#71	29 Sep. 2016	Finance Committee Meeting
#71	29 Sep. 2016	Joint meeting of the two committees of the Interior, Justice and Legal System; Transportation Committee Meeting; Social Welfare and Sanitation and Environment Committee Meeting
#71	11 Oct. 2016	Continue to question the President of the Executive Yuan's Policy Address—continued interrogation
#72	5 Oct. 2016	Interior Committee Meeting; Finance Committee Meeting
#72	5 Oct. 2016	Foreign Affairs and National Defense Committee Meeting
#72	6 Oct. 2016	Foreign Affairs and National Defense Committee Meeting
#72	12 Oct. 2016	Public hearing of the House-wide Committee Meeting

Table 2. Minutes of the 6th meeting of the 2nd session of the 9th Legislative Yuan

No.	Dates	Meetings or Topics
#73	5 Oct. 2016	Social Welfare and Hygiene Environment Committee Meeting; Education and Culture Committee Meeting
#73	14 Oct. 2016	Report matters, continue to inquire about the President's Policy Address—continued interrogation
#73	18 Oct. 2016	Continue to inquire about the President's Policy Address—after the inquiries are answered, the Executive Yuan's reply part and the members' question part
#74	5 Oct. 2016	Transportation Committee Meeting
#74	13 Oct. 2016	House-wide Committee Meeting
#74	17 Oct. 2016	House-wide Committee Meeting
#74	19 Oct. 2016	House-wide Committee Meeting
#74	20 Oct. 2016	House-wide Committee Meeting

¹ The official website is available at: <http://lci.ly.gov.tw/>

The corpus contained 18,050 utterances from 159 speakers, including 395,235 words. Of all the recorded utterances, 1,089 utterances (6%) were marked by the Legislative Yuan as incomplete, which were defined as interrupted sentences (see Section 2.2). The interrupted sentences included 18,629 words, and the interrupting sentences totaled 18,370 words, as shown in Table 3 below. The interrupting sentences were mostly questions and statements, with exclamations making up about 9% of all the interrupting sentences.

Table 3. Sentence types in the corpus

Sentence Types	Sentence Counts	Word Counts
Regular sentences	15,872	358,236
Interrupted sentences	1,089	18,629
Interrupting sentences	1,089	18,370

3.2 Limitations

To clarify the limitations of this study, a few things should be noted. First, the analysis was solely based on the official written transcripts of the parliamentary interpellations. We did not interpret what might or might not count as an interrupted sentence but instead relied fully on the definition provided by the Legislative Yuan.

Second, the Legislative Yuan did not transcribe the conversations according to the conventions of Conversation Analysis. Information about intonation, among other speech elements, was not available. The Legislative Yuan did not provide an official definition of exactly which circumstances stenographers were advised to mark an interruption with an ellipsis. However, based on our extensive reading of the materials, the official transcripts were consistent in terms of formatting and level of transcription detail. For example, throughout the transcripts, final particles expressing emotions, such as *a* (啊), *o* (喔), and so on, occurred frequently, as expected.

Third, we did not examine audio or video recordings to verify that each interruption was accurately recorded to the syllable. We noticed, however, that cut-off words were not found in the transcript, such as *ban...* (辦...) for *banshichu* (辦事處) ‘office’. The microphones of the speakers were open during the conversations; it was, therefore, possible for any speaker to speak over someone else. The stenographers could hear the end of an interrupted sentence as well as the beginning of an interrupting sentence as clearly as anyone else.

Finally, we had no information about overlaps. It was reasonable to assume that some overlapping occurred, but due to the nature of the transcriptions, there was no way of knowing when and how the overlapping occurred.

Despite those limitations, the database represents the first large-scale statistical, and linguistic attempt to look into the phenomenon of interruptions in the Chinese-language context. We maintain that the amount of data extracted allowed us to address how interruptions in the discourse were realized more objectively, repeatedly, and in a data-driven way than using a small-scale but very detail-oriented approach.

3.3 Calculation

3.3.1 Test 1: Word-by-Word Comparison

In order to realize whether a word appeared more or less often in interrupted and interrupting sentences than in regular sentences, we compared words on both the word level and the parts-of-speech level. We compared the frequency of each word against itself across the three sentence types (i.e., interrupted, interrupting, and regular) using a two-side *t*-test based on the weight differences that each word exhibited.² The weight differences were approximately normally distributed (see Figures 1 to 4); therefore, we calculated a z-score³ that represented the standardized deviation from the mean value. Its associated *p*-value indicated how likely it was that the observed deviation would occur due to chance, rather than, in our case, caused by interruption effects. We repeated this calculation independently for each word in the interrupted and interrupting sentences. Table 4 demonstrates this with an example of the word *qishi* (其實) ‘actually’:

Table 4. Word frequency differences across the three different sentence types

Words	English	Weight	Weight Regular [†]	Weight Difference	Z-scores	P-values
Interrupted Sentences						
其實	actually	0.0033	0.0019	0.0013	3.6883	0.0002
Interrupting Sentences						
其實	actually	0.0013	0.0019	-0.0006	-2.1062	0.0352

Note: [†]Weight Regular=weight in the regular sentences.

² The weight of a word refers to the proportion of the sum of all instances of a word x_i over the sum of all in-stances of all other words x_j within the same sentence type, $w=(\sum x_i)/(\sum x_j)$, either interrupted, interrupting, or regular. Weight difference refers to comparing weights between sentence types.

³ The standard z-score is defined as $z=(x-\mu)/\sigma$.

Figures 1 to 4 show the distribution of the weight differences for the words between interrupted, interrupting, and regular sentences. The plots in Figures 1 and 3 show all the words, and the plots in Figures 2 and 4 only show frequently occurring words (count ≥ 10 in the entire transcript).

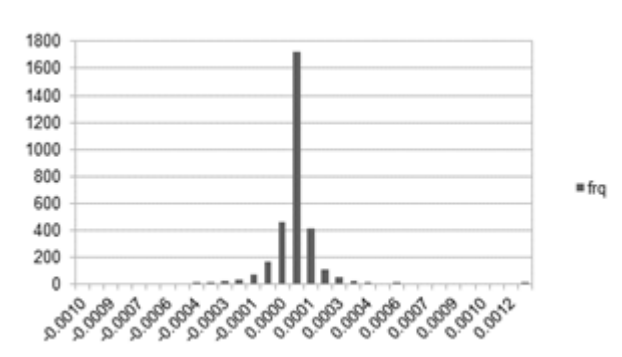


Figure 1. Distribution of weight differences for all words in the interrupted sentences

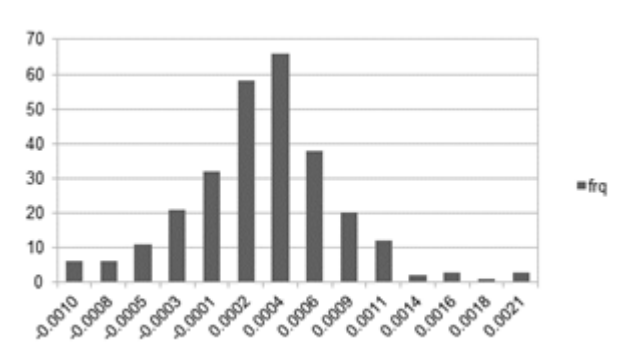


Figure 2. Distribution of weight differences for words with ≥ 10 occurrences in the interrupted sentences

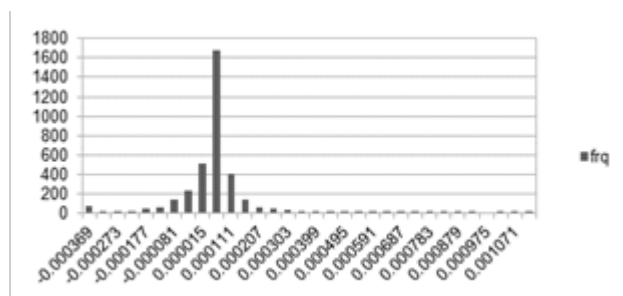


Figure 3. Distribution of weight differences for all words in the interrupting sentences

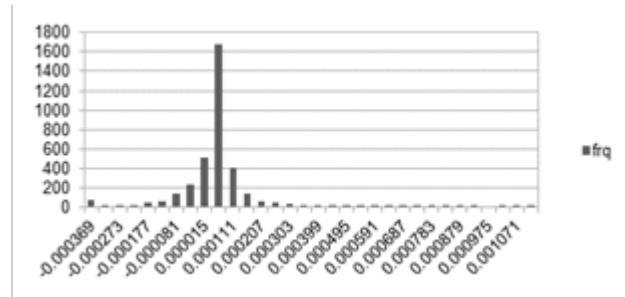


Figure 4. Distribution of weight differences for words with ≥ 10 occurrences in the interrupting sentences

The figures above demonstrate that the weight differences across all (or only a subset of) the words were normally distributed. Therefore, we calculated a z-score for each word, and its associated p -value measured how likely that, compared with its baseline in the regular sentence, an increased or decreased occurrence of a word in the interrupted and interrupting sentences would randomly occur. As the analysis shows, some of the frequency shifts were explained by interruption effects.

3.3.2 Test 2: Comparing Word Rankings in Parts-Of-Speech Categories across Sentence Types

A second test was performed to check the variability of word rankings within the same parts-of-speech category. First, every word was given three rankings in its respective parts-of-speech category⁴ or categories according to its frequencies in interrupted (ED-Ranking), interrupting (ING-Ranking), and regular (Regular Ranking) sentences. We compared the rankings against each other and excluded all words that had a ranking in one sentence type (mostly the regular sentence type) but did not appear in another sentence type (mostly the interrupted and interrupting sentence types). The resulting pairs of rankings were compared by a two-tailed Wilcoxon signed-rank test.⁵ This test measured the significance of the word ranking differences (called “rank shifts”) in each parts-of-speech category compared to itself across the different

⁴ Some Chinese words (character combinations) are associated with more than one word class. For example, words like *zhiqian* (之前) ‘before’ can be temporal nouns (Nd) or locative nouns (Ng); words like *buran* (不然) ‘otherwise, it is not’ can be either conjunctions (Cbb) or stative intransitive verbs (VH). Again, the categorization of each word in each sentence was tagged by the automated sentence tagger provided by CKIP.

⁵ We used the Wilcoxon signed-ranks test calculator available at: <https://www.socscistatistics.com/tests/signedranks/default.aspx>

sentence types.⁶ The test returned a *p*-value that suggested how likely the observed ranking differences in an entire parts-of-speech category would be attributed to random chance.⁷ Our null-hypothesis was that the ranking of words within a category would be similar across sentence types (see Table 5). A preliminary explanation of instances in which this was not the case will be provided in the analysis.

The calculation of rank shifts was undertaken with the following considerations in mind. We limited the scope to only words that had at least 10 or more occurrences in the regular sentence type because, based on the absolute differences in type size, the low-frequency words in the larger-size regular sentence type had a much higher ranking than the low-frequency words in the smaller-size interrupted and interrupting sentence types when those words were included in the smaller-size sentence types. Due to the nature of the Wilcoxon signed-ranks test, these differences added up and caused the compounded overall ranking difference to be very large, resulting in false positives (i.e., very low *p*-values).

Following the same logic, we excluded all words in the regular sentence type that were not included in the interrupted and interrupting sentence types (i.e., zero-entry due to their low frequencies). Furthermore, to reflect the low limit of at least 10 occurrences in the regular sentence type, we also needed to take precautions against low-frequency words in the other two sentence types. Instead of using a hard cut-off as an arbitrary limit, we used the numerically highest ranking number minus 1 as an indicator for how many words above should be included. For example, in the numeral adverbs (Da) class,⁸ the highest ranking number in the interrupted column was 8 (in total, five words had a ranking of 8, all with a frequency of one); therefore, we included only the first seven words. This was a simple yet robust method that, in effect, excluded most words in a category with the lowest frequency (=1), while at the same time it was sensitive to varying category sizes. In only a very few cases did we need to apply a hard cut-off of 100 words, such as when this method failed to limit the total word count to $n=200$.⁹ In some other cases, only a *W*-value could be calculated, but not a *p*-value, because critical *N* ($N \geq 20$) was not reached. In the few cases of a very low *N* size (≈ 10), the Wilcoxon signed-ranks test failed. Due to space limitation, Table 5 shows a list of selected categories.

⁶ Rank shift is a simpler term for significant rank differences between types of sentences calculated by the Wilcoxon signed-rank test.

⁷ In the few cases of *N* being lower than 20 words, a *p*-value could not be calculated so we used the *W*-value instead.

⁸ Examples of the numeral adverbs (Da) class include *dayue* (大約) ‘about’, *zuiduo* (最多) ‘at most’, etc.

⁹ $n=200$ is the maximum count of items generally recommended for a Wilcoxon signed-rank test.

Table 5. Within-category (parts of speech) comparison of word frequency rankings across the interrupted, interrupting, and regular sentence types

Parts-of-speech Categories	P-value Rank Shifts ED↔ING	<i>n</i>	P-value Rank Shifts R↔ED	<i>n</i>	P-value Rank Shifts R↔ING	<i>n</i>
Adjectives						
A (ex. 公共)	0.3953	22	0.1031	20	0.4533	24
Conjunctions						
Caa <i>coordinating</i> (ex. 和)	not significant ¹⁰	12	not significant ¹¹	10	not significant ¹²	12
Cbb <i>subordinating</i> (ex. 如果)	0.6171	62	0.2627	41	0.5029	30
Adverbs						
D (ex. 儘量)	0.7490	100	0.1310	130	0.2670	123
Dfa <i>with degree</i> (ex. 非常)	0.7263	20	0.9283	14	not significant ¹³	14
Nouns						
Na <i>regular</i> (ex. 問題)	0.2891	100	**0.0017	100	0.0629	100
NC <i>place names</i> (ex. 大陸)	0.4902	100	0.2041	96	0.3077	95
Ncd <i>locative</i> (ex. 裡面)	0.9761	20	0.1556	18	0.1310	20
Nd <i>temporal</i> (ex. 目前)	0.5093	31	0.4777	19	0.2713	31
Neqa <i>count nouns</i> (ex. 某些)	0.7949	28	0.7114	25	0.2041	28
Nh <i>pronouns</i> (ex. 我)	0.6384	20	0.5029	19	*0.0414	18
Prepositions						
P (ex. 至於)	0.3271	50	0.3681	53	0.9362	45
Verbs						
VA <i>intransitive</i> (ex. 犯罪)	0.5353	42	0.7566	38	0.0873	44
VC <i>transitive</i> (ex. 提出)	0.0000	186	0.1499	162	0.1096	177
VD <i>ditransitive</i> (ex. 提供)	0.5093	11	data insufficient	7	*0.0340	15
VE <i>verb + subclause</i> (ex. 認為)	0.3173	76	*0.0357	72	0.8026	65
VF <i>verb + verbal phrase</i> (ex.)	0.6101	14	0.4237	11	0.7566	12

¹⁰ W=16, critical value for W at N=8 ($p<0.05$) is 3.

¹¹ W=17, critical value for W at N=9 ($p<0.05$) is 5.

¹² W=13, critical value for W at N=9 ($p<0.05$) is 5.

¹³ W=10, critical value for W at N=6 ($p<0.05$) is 0.

Parts-of-speech Categories	<i>P</i> -value Rank Shifts ED↔ING	<i>n</i>	<i>P</i> -value Rank Shifts R↔ED	<i>n</i>	<i>P</i> -value Rank Shifts R↔ING	<i>n</i>
繼續)						
VG <i>category</i> (ex. 成為)	0.8493	23	0.0836	20	0.3222	21
VH <i>stative intransitive</i> (ex. 努力)	0.3953	127	0.3271	113	0.1615	116
VHC <i>causative</i> (ex. 落實)	not significant ¹⁴	9	0.8887	16	0.5353	12
VJ <i>stative transitive</i> (ex. 沒有)	0.3030	47	0.5823	36	0.4237	40
VK <i>stative + subclause</i> (ex. 希望)	0.8729	36	0.0989	35	0.7490	28

Note: ED=interrupted sentence type; ING=interrupting sentence type; R=regular sentence type; *= $p < 0.05$; **= $p < 0.01$.

As Table 5 demonstrates, most categories had comparable internal word rankings across the different sentence types, as expected. First, this suggested that not all words within a parts-of-speech category participated in the phenomenon of interruptions. Put differently, interruptions were not predominantly caused at the parts-of-speech level. However, we have to be careful with this statement because our analysis in Section 4 will also show that conjunctions, adverbs, and pronouns were overrepresented in the keywords, and nouns, verbs, and adjectives were underrepresented, compared with the overall percentage in the lexicon and the 100 most common words in the corpus. On the question of the relationship between parts of speech and interruptions, the status and function of a keyword was not predicated on its specific placement within a parts-of-speech category. Yet some parts-of-speech categories participated more actively in interruptions than other categories did. The underlying mechanism of interruptions might be best understood as being related to both frequency and semantics. On the one hand, most of the keywords were very common words, while on the other hand, we did not find that the keywords were semantically random; rather, we found that there was a relationship between discourse functions and strategies, as will be shown in Section 5.

3.4 Word Tagging and Parts of Speech

All the sentences in this corpus were tagged automatically, with no human intervention or correction, using the CKIP sentence tagger developed by the CKIP Lab at Academia Sinica (Ma & Chen, 2003).¹⁵ CKIP differentiates 46 parts-of-speech categories, organized in 10 main groups.¹⁶ Automated tagging arguably includes mistakes and is different from a parts-of-speech

¹⁴ $W=17$, critical value for W at $N=8$ ($p < 0.05$) is 3.

¹⁵ See ckip.iis.sinica.edu.tw

¹⁶ See ckipsvr.iis.sinica.edu.tw

analysis provided by a native speaker. We always followed the CKIP system because automated word tagging creates repeatable and comparable results.

3.5 Defining Keywords

Two conditions had to be satisfied to be classified as a keyword. First, a keyword was a word that appeared at least 10 times in the entire corpus, and second, its associated p -values (one for its frequency in interrupted sentences and one for its frequency in interrupting sentences) had to be at least 0.05 or lower. The p -value was interpreted as the likelihood that the difference in frequency between its usage in regular sentences and interruption sentences would occur due to random chance. This threshold was set arbitrarily but was based on the assumption that often-appearing interruption effects would be observed by often-appearing structural features. We did not exclude the possibility that some other systematic trigger stimulus also existed.

We referred to significant words as keywords in the study, whether they appeared in interrupted or interrupting sentences. A strict distinction between these two sentence types (i.e., interrupted and interrupting) was unnecessary because each type was clearly delineated whenever they appeared.

3.5.1 Position of Keywords in a Sentence

Once a keyword was statistically identified, we did not know where exactly it appeared in a sentence. As explained above, we did not calculate the distance from the truncated turn-final position because in considering word position, we did not know how to relate “difference from turn-final” to any random position in a regular sentence. We considered the position of a word in a regular sentence to be random and to co-vary with content-dependent factors. Therefore, a keyword was accounted for only by its appearance in an interrupted, interrupting, or regular sentence, not by its location. As a consequence, any keyword(s) in a sentence—by itself or in cooperation with other keywords—was regarded as important, independent of location.

3.5.2 Marking Keywords in Example Sentences

In each given example sentence in the analysis, the selection of the keyword was used for exemplary purposes only. For example, our analysis showed that pronouns were important in explaining interruptions. Therefore, to demonstrate the importance of pronouns, we selected a few example sentences from the database that included a pronoun. According to our subjective reading, these example sentences clearly demonstrated the interruption effect of pronouns, as shown in (1) below¹⁷:

¹⁷ Note that in the example sentences shown, specific selected keywords were not necessarily responsible for triggering the interruptions. The example sentences were meant to demonstrate that a specific

- (1) 195:a 我跟李部長也有在討論這個問題…
‘Minister Li and I are also discussing this issue…’

Note that in (1) above, the word *wenti* (問題) ‘problem’ was not considered a keyword because its associated *p*-value was not significant. Although it appeared very often, its frequency was relatively consistent across all types of sentences—regular, interrupted, and interrupting. The same was the case for all the other words in this example sentence, including *gen* (跟) ‘with/and’, *ye* (也) ‘also, too’, *you* (有) ‘have’, *zai* (在) ‘is, in, (grammatical particle)’, *taolun* (討論) ‘discuss’, *zhe* (這) ‘this’, and *ge* (個) ‘(counting particle)’. The only other possible valid keyword, next to *wo* (我) ‘I’, according to the test results, was *buzhang* (部長) ‘minister’, as discussed in the names and personal titles section. We chose *wo* ‘I’ here on subjective grounds. In fact, *wo* ‘I’ and *buzhang* ‘minister’ might have co-triggered the interruption as two or more keywords can jointly trigger interruptions.

4. Analysis

This section will present the analysis of the interrupted and interrupting sentences, which were organized by parts of speech. The most common keywords were not distributed in the same way across parts-of-speech categories as suggested by their overall number in the corpus (and by extension in the lexicon). Conjunction, adverb, and pronoun keywords were overrepresented, while noun, verb, and adjective keywords were underrepresented. This suggests that, if anything, some parts-of-speech categories were more important for interruptions than others and, conversely, that some parts-of-speech categories play a less important role. Table 6 shows the comparison of the distribution of total word counts in the corpus in each parts-of-speech category and their respective percentages with that of the total word counts of the keywords. In both cases, only word types were calculated.

keyword was *somehow* involved in the speech act of interruption, but this did not imply that it alone caused the interruption. In fact, it was possible that all the keywords identified in this study only represented an epiphenomenal linguistic pattern, one that was only correlated with interruptions on the surface level but was unrelated to causation.

Table 6. Distribution of parts-of-speech categories in the corpus vs. distribution of keywords

Parts of Speech	Word Counts (Corpus)	% of Total	Word Counts (Keywords)	% of Total
(A) Adjectives	265	1.5%	0	0.0%
(C) Conjunctions	130	0.7%	9	9.1%
(D) Adverbs	813	4.6%	22	22.2%
(N) Nouns	9,017	51.0%	29	29.3%
(Nh) Pronouns	57	0.3%	10	10.1%
(O) Others	186	1.1%	11	11.1%
(V) Verbs	7,205	40.8%	18	18.2%
Total	17,673	100.0%	99	100.0%

Furthermore, keywords were not to be confused with the most frequent words. A keyword is different from a regular frequent word in that it shows a specific significant frequency shift across different sentence patterns, whereas a regular frequent word appeared similarly frequently across all sentence types. We demonstrated this by looking at how many of the most common words also appeared in the list of keywords. Keywords that were also among the most common words were arguably less strictly related to interruptions than keywords only.

Table 7. Comparison of the top 100 most common words that also appeared in the list of keywords

Parts of Speech	Word Counts (Most Common 100 Words)	Most Common Words – Keywords	Most Common Words – Keywords	Keywords – Most Common Words	Keywords – Most Common Words
(A) Adjectives	0	0	0	0	
(C) Conjunctions	11	8	3	1	不過
(D) Adverbs	22	17	5	5	一定, 不要, 又, 比較, 當然
(N) Nouns	26	18	8	11	一些, 上, 中, 主席, 事實, 基本, 委員會, 政策, 期, 案件, 話
(Nh) Pronouns	10	10	0	0	
(O) Others	13	8	5	3	嘛, 得, 至於
(V) Verbs	18	12	6	6	問, 回答, 有關, 為, 謝謝, 進行
Total	100	73	27	26	

Conversely, words that were the most common across any sentence type but did not appear in the list of keywords were least relevant for the topic at hand.

As Table 7 demonstrates, the most frequent words and keywords were notionally different. To give a general picture, Table 7 does not differentiate between keywords occurring in either interrupted or interrupting sentences—or its increased or decreased tendencies for that matter—but reports only the overall sum. If a certain keyword appeared in both the interrupted and interrupting sentences in a significant way, it was counted only once in the table. However, as we proceed with the analysis in more detail below, we count keywords in the interrupted and interrupting sentences separately.

Table 7 also shows that although some keywords were also the most frequent, roughly one-quarter (26 out of 99) of the keywords were not. Conversely, 27 words were very frequent but did not count as keywords. As a general rule, words with significant frequency shifts in the interrupted sentences were non-significant in the interrupting sentences, and vice versa. This is important to note when comparing numbers between the sentence types. In the following, we will introduce each parts-of-speech category and its contribution to interruptions in detail before we discuss the possible relationships between the parts of speech and discourse functions and strategies in Section 5.

4.1 Pronouns

In terms of the interrupted sentences, we observed that first-person pronouns were used significantly more often than in the regular sentences, but the use of second-person pronouns significantly decreased, as shown in Table 8 below. We counted 766 instances of *wo* (我) ‘I’, *women* (我們) ‘we’, and *ziji* (自己) ‘self’—in contrast to only 106 instances of *ni* (你) ‘you’ (sg.), *nimen* (你們) ‘you’ (pl.), and *nin* (您) ‘you’ (polite)—in the interrupted sentences. This suggests a situation in which the interrupted speaker adopted a defensive strategy to explain his or her view while being constantly attacked. It could also mean that the legislators were more likely to be interrupted when they spoke about themselves and their in-group.

Table 8. Pronouns in the interrupted sentences

Parts of Speech	Words	English	Weight Interrupted*	Weight Regular†	Tendency	P-value
Pronoun (Nh)	我	I, me	0.0229	0.0146	increased	0.0000
Pronoun (Nh)	我們	we	0.0172	0.0112	increased	0.0000
Pronoun (Nh)	本席	myself	0.0006	0.0022	decreased	0.0000
Pronoun (Nh)	你	you	0.0036	0.0088	decreased	0.0000
Pronoun (Nh)	你們	you (pl.)	0.0012	0.0026	decreased	0.0000
Pronoun (Nh)	您	you (polite)	0.0009	0.0022	decreased	0.0000

Note: *Weight Interrupted=weight in the interrupted sentences; †Weight Regular=weight in the regular sentences.

Table 8 above shows a sample of pronouns with significant frequency shifts between the interrupted and the regular sentences. Out of the 24 pronouns found in the interrupted sentences, only eight (six shown here) exhibited significant frequency shifts.

In the interrupting sentences, on the other hand, we observed a significant increase in both second-person (singular and plural) pronouns and first-person (singular) pronouns, as shown in Table 9 below. This suggests that, in contrast to the defensive discourse strategy in the interrupted sentences above, the interrupters often directed their verbal attacks toward an individual or a group.

Table 9. Pronouns in the interrupting sentences

Parts of Speech	Words	English	Weight Interrupting*	Weight Regular†	Tendency	P-value
Pronoun (Nh)	你	you	0.0209	0.0088	increased	0.0000
Pronoun (Nh)	我	I, me	0.0183	0.0146	increased	0.0000
Pronoun (Nh)	你們	you (pl.)	0.0052	0.0026	increased	0.0000
Pronoun (Nh)	他們	they, them	0.0038	0.0028	increased	0.0020
Pronoun (Nh)	我們	we	0.0089	0.0112	decreased	0.0000

Note: *Weight Interrupting=weight in the interrupting sentences; †Weight Regular=weight in the regular sentences.

Table 9 shows a sample of pronouns with significant rank shifts between the interrupting and the regular sentences. Many more words, marked as pronouns by CKIP, did not follow this pattern, for example: *nin* (您) ‘you (polite)’, *benxi* (本席) ‘myself’, *dajia* (大家) ‘everybody’, *shei* (誰) ‘who’, and *benshen* (本身) ‘myself’, among others. Out of the 28 pronouns found in the sentences, only seven (five shown here) had a *p*-value <0.05 in the interrupting sentences.

4.2 Conjunctions

As a parts-of-speech category, conjunctions did not exhibit sufficiently different occurrence patterns between the interrupted, interrupting, and regular sentences. However, some individual conjunctions showed significant interruption effects, specifically *yinwei* (因為) ‘because’, *suoyi* (所以) ‘therefore’, *danshi* (但是) ‘but’, and *ruguo* (如果) ‘if’. We observed this tendency also for the noun *yuanyin* (原因) ‘reason’, but to a lesser extent. These functional words were used to indicate reasoning and to present arguments in a logical fashion. This suggests that the interlocutors were more likely to be interrupted when they used reason during political debates.

In the interrupting sentences, the conjunction *suoyi* ‘therefore’ followed the pattern of *yinwei* ‘because’; however, whereas *yinwei* ‘because’ showed a significant increase in usage only in the interrupted sentences, *suoyi* ‘therefore’ was used significantly more often in both types of sentences as a general marker of interruptions. Break points were associated with the

use of restrictive conjunctions such as *danshi* ‘but’ and *ruguo* ‘if’ and to some degree also with *keshi* (可是) ‘but’. When used by the interrupted speaker, restrictive conjunctions indicated the speaker’s intention to introduce a different argument. From the opponent’s perspective, however, formulating a nuanced argument invited opposition. Out of the 55 conjunctions included in the list of interrupted sentences, only the six listed below in Table 10 showed significant rank shifts:

Table 10. Conjunctions in the interrupted sentences

Parts of Speech	Words	English	Weight Interrupted*	Weight Regular†	Tendency	P-value
Conjunct. (Cbb)	因為	because	0.0084	0.0040	increased	0.0000
Conjunct. (Cbb)	所以	therefore	0.0063	0.0043	increased	0.0000
Conjunct. (Cbb)	如果	if	0.0055	0.0035	increased	0.0000
Conjunct. (Cbb)	但是	but	0.0040	0.0029	increased	0.0026
Conjunct. (Cbb)	不過	but	0.0013	0.0006	increased	0.0434
Conjunct. (Caa)	及	and	0.0012	0.0022	decreased	0.0054

Note: *Weight Interrupted=weight in the interrupted sentences; †Weight Regular=weight in the regular sentences.

In the interrupting sentences, the situation was different, as shown in Table 11. Only the conjunction *suoyi* ‘therefore’ showed a significant increase in usage frequency in the interrupting sentences, arguably because *suoyi* ‘therefore’ was often used to ask questions and formulate a conclusion in a dialogue. The remaining four words were not associated with semantic patterns and hence remain currently unexplained. Out of the 44 conjunctions included in the list of interrupting sentences, the five conjunctions below in Table 11 showed significant rank shifts:

Table 11. Conjunctions in the interrupting sentences

Parts of Speech	Words	English	Weight Interrupting*	Weight Regular†	Tendency	P-value
Conjunct. (Cbb)	所以	therefore	0.0064	0.0043	increased	0.0000
Conjunct. (Caa)	及	and	0.0010	0.0022	decreased	0.0002
Conjunct. (Caa)	和	and	0.0011	0.0019	decreased	0.0071
Conjunct. (Caa)	與	and	0.0010	0.0017	decreased	0.0186
Conjunct. (Cbb)	而	and, but	0.0015	0.0022	decreased	0.0217

Note: *Weight Interrupting=weight in the interrupting sentences; †Weight Regular=weight in the regular sentences.

4.3 Nouns

Nouns comprised the largest parts-of-speech category, making up roughly 50% of the corpus. However, they contributed only 29 words to the list of keywords, which was 29% of the keywords. Of those 29 keywords, 11 nouns were in the list of keywords but were not found in the top 100 most common words, so they were the best candidates for further analysis. Tables 12 and 13 below show some of the nouns of interest:

Table 12. Noun keywords in the interrupted sentences that were not also in the top 100 most common words

Parts of Speech	Words	English	Weight Interrupted*	Weight Regular [†]	Tendency	<i>P</i> -value
Nouns (Na)	事實	fact	0.0017	0.0008	increased	0.0107
Nouns (Na)	基本	basis	0.0013	0.0004	increased	0.0195
Nouns (Nf)	期	term	0.0013	0.0002	increased	0.0043
Nouns (Na)	案件	case	0.0014	0.0007	increased	0.0319

Note: *Weight Interrupted=weight in the interrupted sentences; [†]Weight Regular=weight in the regular sentences.

Table 13. Noun keywords in the interrupting sentences that were not also in the top 100 most common words

Parts of Speech	Words	English	Weight Interrupting*	Weight Regular [†]	Tendency	<i>P</i> -value
Nouns (Na)	話	talk, speech	0.0025	0.0012	increased	0.0001
Nouns (Na)	主席	chairman	0.0002	0.0013	decreased	0.0002
Nouns (Na)	政策	policy	0.0005	0.0013	decreased	0.0063
Nouns (Nc)	委員會	committee	0.0001	0.0008	decreased	0.0170

Note: *Weight Interrupting=weight in the interrupting sentences; [†]Weight Regular=weight in the regular sentences.

Furthermore, referring back to Table 5, regular nouns (Na), as a category, showed one of the most significant rank shifts between the interrupted and regular sentences ($p < 0.0017$). The statistical explanation was that regular nouns—as well as regular adverbs (D) and verbs (VE) (to a lesser degree)—in the interrupted sentences yielded the highest count of absolute ranking differences: 6% for the top 100 words, followed by 26% for the top 1,000 and 32% for the top 10,000 words. This suggested that frequently used nouns in the regular sentences did not appear in the interrupted sentence, and vice versa. The discourse explanation was that the regular nouns were highly content-dependent and were supposed to co-vary strongly with the topic.

Another issue involved temporal nouns. The CKIP team classified the words *muqian* (目前) ‘currently’, *guoqu* (過去) ‘past’, and *dangshi* (當時) ‘now’ as temporal nouns in Chinese because they either were composed of a noun (e.g., *mu* ‘eye’, *shi* ‘time’) or could be used as a noun (e.g., *guoqu* ‘past’). In English, however, they are not classified as nouns but as adverbials. In the interrupted sentences, these three words showed a significant shift in frequency. We argue that they functioned as downtoners that were used in defensive discourse strategies, indicating a temporal limitation to what was being said to appear restrained or cautious.

Table 14. Temporal nouns in the interrupted sentences

Parts of Speech	Words	English	Weight Interrupted*	Weight Regular†	Tendency	P-value
Temp. Noun (Nd)	目前	currently	0.0035	0.0018	increased	0.0000
Temp. Noun (Nd)	過去	ago, before, previously	0.0017	0.0012	increased	0.1862
Temp. Noun (Nd)	當時	before	0.0012	0.0006	increased	0.1614
Temp. Noun (Nd)	未來	in future	0.0007	0.0012	decreased	0.0902

Note: *Weight Interrupted=weight in the interrupted sentences; †Weight Regular=weight in the regular sentences.

Although the word *muqian* ‘currently’ was the only temporal noun that showed significant rank shifts in the interrupted sentences (see Table 14), we nonetheless argue that the word was a good example of a downtoner, a keyword that appeared significantly more often in the sentences that were being interrupted. Our argument is that it represented more than just a temporal downtoner and reflected a certain attitude of what might be called “pseudo-objectivity,” that is, to appear objective and scientific. As such, it created argumentative boundaries that were very likely to be challenged by the opponent. In the interrupting sentences, *muqian* ‘currently’ exhibited a contrastive tendency and was used much less as an aggressive strategy. This suggested that *muqian* ‘currently’ played an important role in inviting interruptions, albeit arguably not intended by the speaker. In terms of the interrupting sentences, the word *xianzai* (現在) ‘now’ featured a higher frequency in the interrupting sentences than in the regular sentences, as shown in Table 15 below:

Table 15. Temporal nouns in the interrupting sentences

Parts of Speech	Words	English	Weight Interrupting*	Weight Regular†	Tendency	P-value
Temp. Noun (Nd)	現在	now	0.0060	0.0040	increased	0.0000
Temp. Noun (Nd)	目前	currently	0.0012	0.0018	decreased	0.0451

Note: *Weight Interrupting=weight in the interrupting sentences; †Weight Regular=weight in the regular sentences.

4.4 Adjectives

As a parts-of-speech category, the rank shifts of adjectives between the interrupted and the regular sentences showed a weak tendency ($p \approx 0.10$). The adjectives *yiding* (一定) ‘necessary’, *yiban* (一般) ‘regular’, and *jiben* (基本) ‘basic’ exhibited weak interruption effects but not at a significant level ($p > 0.05$). The reason might have been that the adjectives belonged to a group of content words and were expected to co-vary with speech content more than with interruption patterns.

4.5 Verbs

In general, verb keywords were underrepresented in both lists of the top 100 most common words and significant keywords. Verbs made up roughly 41% of all the words in the database but accounted for only 18% of the keywords. Furthermore, there were 18 words in the top 100 most common words. This suggested that verbs were less relevant as keywords. In total, three subtypes of verbs showed a significant rank shift: transitive verbs (VC) were significantly different between the interrupted and the interrupting sentences ($p < 0.0455$); ditransitive verbs (VD) were different between the interrupting and the regular sentences ($p < 0.0340$); and verbs with adjunct subclauses (VE) were different between the interrupted and the regular sentences ($p < 0.0357$), as shown in Table 16 below:

Table 16. Verb classes with significantly different within-rank shifts between various categories

Parts of Speech	<i>P</i> -value Rank Shift ED↔ING	<i>n</i>	<i>P</i> -value Rank Shift R↔ED	<i>n</i>	<i>P</i> -value Rank Shift R↔ING	<i>n</i>
VC transitive (ex. 提出)	*0.0455	186	0.1499	162	0.1096	177
VD ditransitive (ex. 提供)	0.5093	11	data insufficient	7	*0.0340	15
VE verb + subclause (ex. 認為)	0.3173	76	*0.0357	72	0.8026	65

Note: ED=interrupted sentences; ING=interrupting sentences; R=regular sentences; *= $p < 0.05$.

The question of why VC and VD verbs showed significant interruption effects remains unanswered at the moment. In terms of the VE verbs, we also speculate that they showed significance because they often appeared at the beginning of a sentence before the interruption took place. The 10 most common VE verbs were: *qingwen* (請問) ‘may I ask’ ($p < 0.0000$), *jiang* (講) ‘say’ ($p < 0.0210$), *tidao* (提到) ‘mention’ ($p < 0.0531$), *dafu* (答覆) ‘to answer’ ($p < 0.0572$), *zhixun* (質詢) ‘to question’ ($p < 0.0808$), *qingjiao* (請教) ‘to consult’, *wen* (問) ‘ask’, *xunda* (詢答) ‘inquire’, *renwei* (認為) ‘argue’, and *kandao* (看到) ‘have seen’, all of which belonged to a subclass of discourse markers that were used to talk about speech content. The first two words showed interruption effects. *Qingwen* ‘may I ask’ was used significantly less often in the regular

sentences; *jiang* ‘say’, on the other hand, was used significantly more. The verbs *dafu* ‘to answer’ and *zhixun* ‘to question’ were used more by the questioning opponents and much less by the queried persons.

The most common verbs with a significantly increased occurrence in the interrupted sentences were *you* (有) ‘there is, have’, *meiyou* (沒有) ‘do not have’, *jiang* (講) ‘say’, *xiwang* (希望) ‘to hope’, and *kan* (看) ‘see’, and those with a significantly decreased occurrence in the interrupted sentences included *qing* (請) ‘please’, *rang* (讓) ‘let’, and *xiexie* (謝謝) ‘thank you’. With the exception of *xiexie* ‘thank you’, all of these words were found among the top 100 most common words in the corpus. Table 17 below shows some verbs with significantly changed behavior between the interrupted and the regular sentences. Most of these verbs were related to speech acts, which additionally explains why they appeared in the interrupted sentences more often.

Table 17. Verbs in the interrupted sentences

Parts of Speech	Words	English	Weight Interrupted*	Weight Regular†	Tendency	P-value
Verb (V_2)	有	have	0.0179	0.0140	increased	0.0000
Verb (VJ)	沒有	do not have	0.0041	0.0030	increased	0.0025
Verb (VE)	講	talk	0.0028	0.0019	increased	0.0210
Verb (VK)	希望	hope	0.0031	0.0023	increased	0.0277
Verb (VE)	請問	may I ask	0.0001	0.0017	decreased	0.0000
Verb (VF)	請	please	0.0017	0.0027	decreased	0.0000
Verb (VJ)	謝謝	thank you	0.0005	0.0013	decreased	0.0245

Note: *Weight Interrupted=weight in the interrupted sentences; †Weight Regular=weight in the regular sentences.

Verb keywords in the interrupting sentences were mostly discourse-relevant verbs, a commonality they shared with the interrupted sentences. They often functioned as pragmatic markers or as short replies and were placed at the beginning of the interrupting sentence. Take, for example, *dui* (對) ‘correct, yes’ in (2) below:

- (2) 1122:a 第一個，如果我們能不用核一廠、核二廠，就儘量不用…
 ‘The first one, if we can avoid the first nuclear plant and the second nuclear plant, try not to use it as much as possible…’
- 1122:b 對，大家都歡迎。
 ‘Yes, everyone is welcome.’

We are aware that in Chinese grammar the negator *meiyou* (沒有) ‘do not have’ in (3) below is regarded as an adverb when used to negate an event/activity. CKIP, however, classifies it as a stative intransitive verb on these occasions. This might have also been the case for similar instances, as we always followed the CKIP classification in this study.

- (3) 8024:a 如果我剛才的回答有讓...
 ‘If my answer just now makes...’
 → 8024:b 沒有，你沒有回答，那是陳超明自己在自問自答，你沒有回答，這個還好。
 ‘No, you didn’t answer, that was Chen Chaoming’s own question and answer, you didn’t answer, this is okay.’

Table 18 below shows some (10 out of 18) of the most frequent verbs with significant weight shifts between the interrupting and the regular sentences:

Table 18. Verbs in the interrupting sentences

Parts of Speech	Words	English	Weight Interrupting*	Weight Regular†	Tendency	P-value
Verb (VE)	說	say	0.0049	0.0033	increased	0.0000
Verb (VJ)	沒有	do not have	0.0045	0.0030	increased	0.0000
Verb (VE)	講	talk	0.0039	0.0019	increased	0.0000
Verb (VH)	對	correct, yes	0.0029	0.0018	increased	0.0016
Verb (VK)	知道	know	0.0028	0.0018	increased	0.0021
Verb (VE)	問	ask	0.0015	0.0008	increased	0.0217
Verb (VE)	回答	answer	0.0013	0.0004	increased	0.0094
Verb (VF)	請	please	0.0019	0.0027	decreased	0.0093
Verb (VC)	進行	conduct	0.0002	0.0010	decreased	0.0102
Verb (VJ)	謝謝	thank you	0.0007	0.0013	decreased	0.0372

Note: *Weight Interrupting=weight in the interrupting sentences; †Weight Regular=weight in the regular sentences.

4.6 Adverbs

As a parts-of-speech category, adverbs showed some tendentious interruption effects, though not at a significant level. However, adverbs played a crucial role in the interruptions. They were overrepresented in both lists of the top 100 most common words and significant keywords. Adverbs made up less than 5% of all the words in the corpus, but they represented 22% of the top 100 most common words and 22% of the significant keywords.

We attributed their importance to the fact that adverbs, in general, are often used as modulators to express evaluations or colorize a statement. Looking at the meanings of the often re-occurring adverb keywords in the interrupted sentences, we can explain some of these in light of defensive discourse strategies, which restricted the scope of an argument (e.g., *bijiao* (比較) ‘(comparative)’, *keneng* (可能) ‘possibly’), showed confidence (e.g., *yiding* (一定) ‘necessary’, *dangran* (當然) ‘of course’), or colorized a statement (e.g., *qishi* (其實) ‘actually’.. Also noteworthy are the adverbs *yiding* (一定) ‘definitely’, *you* (又) ‘again’, *bijiao* (比較) ‘(comparative)’, and *dangran* (當然) ‘of course’ because these four keywords were not in the top 100 most common words. Table 19 below shows some significant adverbs in the interrupted sentences:

Table 19. Adverbs in the interrupted sentences

Parts of Speech	Words	English	Weight Interrupted*	Weight Regular†	Tendency	P-value
Adverb (D)	會	will, about to	0.0073	0.0058	increased	0.0000
Adverb (D)	其實	actually	0.0033	0.0019	increased	0.0002
Adverb (D)	已經	already	0.0032	0.0023	increased	0.0171
Adverb (D)	還是	maybe, or	0.0028	0.0019	increased	0.0107
Adverb (D)	可能	possibly	0.0024	0.0016	increased	0.0238
Adverb (D)	當然	of course	0.0021	0.0012	increased	0.0150
Adverb (D)	一定	definitely	0.0020	0.0011	increased	0.0076
Adverb (Dfa)	比較	more (degree)	0.0020	0.0012	increased	0.0324

Note: *Weight Interrupted=weight in the interrupted sentences; †Weight Regular=weight in the regular sentences.

Adverbs also played an important role in the interrupting sentences but to a slightly lesser degree. The adverb *buyao* (不要) ‘do not’ was the only adverb with a significantly increased occurrence in the interrupting sentences that was not found among the top 100 most common words. That a negation adverb occupied such an important position supported the argument that the interrupting party used an offensive discourse strategy. Other common and significant

changes concerned adverbs such as *bu* (不) ‘no’, *ye* (也) ‘also’, *lai* (來) ‘come’, *jiu* (就) ‘(grammatical particle)’, *dou* (都) ‘all’, *yinggai* (應該) ‘should’, *keyi* (可以) ‘can’, *meiyou* (沒有) ‘no, none’, and *hai* (還) ‘still’ (see Table 20).

Table 20. Adverbs in the interrupting sentences

Parts of Speech	Words	English	Weight Interrupting*	Weight Regular†	Tendency	P-value
Adverb (D)	不	no, not	0.0162	0.0107	increased	0.0000
Adverb (D)	就	(particle)	0.0083	0.0060	increased	0.0000
Adverb (D)	會	will	0.0076	0.0058	increased	0.0000
Adverb (D)	沒有	no, none	0.0041	0.0025	increased	0.0000
Adverb (D)	已經	already	0.0034	0.0023	increased	0.0006
Adverb (D)	還	still	0.0032	0.0025	increased	0.0420
Adverb (D)	不要	do not	0.0021	0.0008	increased	0.0001
Adverb (D)	可以	can	0.0050	0.0040	increased	0.0055
Adverb (D)	應該	should	0.0023	0.0031	decreased	0.0106
Adverb (D)	其實	actually	0.0013	0.0019	decreased	0.0352

Note: *Weight Interrupting=weight in the interrupting sentences; †Weight Regular=weight in the regular sentences.

In this section, we analyzed the frequency shifts of keywords in both the interrupted and the interrupting sentences, respectively. We discussed their relationship with the parts-of-speech categories and briefly addressed why the semantics of these words were related to their status as keywords. We noticed that the keywords were very common words with significant frequency shifts across sentence types, and therefore they were different from the regular high-frequency words. Moreover, the keywords were found in almost all the parts-of-speech categories, with the exception of adjectives. However, conjunctions, adverbs, and pronouns stood out for being overrepresented, compared with their footprint in the overall lexicon. That even grammatical particles were considered keywords might sound surprising. Yet, due to their significant frequency shifts and their underlying semantic commonalities, we had to consider this possibility seriously. In the following, we will discuss to what extent the meanings of the keywords were related to offensive and defensive discourse strategies.

5. Discussion

Based on the statistical analysis presented in the previous section, this section will discuss the possible discourse functions of the words that showed a significant increase or decrease in their usage in the interrupted and interrupting sentences compared with that in the regular sentences.

To begin with, the frequency shifts of conjunctions, such as *yinwei* ‘because’ and *suoyi* ‘therefore’, and adverbs, such as *buyao* ‘do not’ and *qishi* ‘actually’, suggests that the interruptions did not happen randomly. Rather, it is reasonable to assume that the changed frequency of the words’ functions was partially related to their semantics. Semantics might also explain why some words signaled specific discourse functions, such as showing disrespect, avoiding a concrete answer, and being downtoners (i.e., expressing pseudo-objectivity), among many more.

The link between individual word meanings and the general incentive structure may rest with discourse functions, which linked the underlying incentive structure to the keywords. The lawmakers often followed an offensive discourse strategy when the opposing ministers adopted a defensive discourse strategy. Generally speaking, an offensive discourse strategy was more associated with interrupting sentences, while a defensive discourse strategy was associated with interrupted sentences.

At this point, it is important to explain how we identified the discourse functions. After all the keywords were statistically identified and linked to offensive (interrupting sentences) and defensive (interrupted sentences) discourse strategies, we then grouped the keywords according to commonly shared themes, semantic fields, and objectives. For example, *buyao* + *V* ‘do not’ + Verb was an often re-occurring pattern in the interrupting sentences, and as such, its theme or objective was associated with ‘negation’, stopping the opponent verbally due to a strong disagreement. Moreover, *wo* ‘I, me’ appeared significantly more often in the interrupted sentences, which was associated with the function of self-reference. In other words, discourse functions provided the argumentative linkage between the semantic fields of the keywords and the objectively observed bifurcation of offensive and defensive discourse strategies under the incentive structure.

A defensive discourse strategy was often applied to avoid political mistakes using downtoners, especially to express pseudo-objectivity in order to appear knowledgeable, objective, and scientific. The following defensive discourse strategies were common: (i) self-reference—to interrupt people when they talked more about themselves than about the subject; (ii) reasonable presentation—to prevent a clear presentation of reason in order to disguise the arguments of the other side or to disallow them to present their case clearly and logically; (iii) over-limitation and pseudo-objectivism—to overly use semantic limiters (e.g., ‘to some degree’, ‘possibly’, ‘in fact’, ‘actually’, etc.); (iv) over-confidence—to use superlatives, amplifiers, and

intensifiers; and (v) subjective or personal evaluation—to express a subjective or personal evaluation.

Offensive discourse functions included the following: (i) negation—to show that the opponent was not true, not right, not informed, or not fit for the job; (ii) adversatives and opposition—to express rejection or contrast of opinion; (iii) superlatives—to use hyperbole to point out extremes, to draw a radical mental image, or to contrast an opponent’s ambiguous statement with an extreme counterpart; (iv) questions—to request more detailed information, (rhetorical questions) to indicate mocking or disbelief, or to request confirmation; and (v) direct address—the interrupter) to directly address his or her opponent either by name, position, or personal pronoun.

In what follows, we will discuss the interrupted sentences (defensive discourse strategies) before we look at the interrupting sentences (offensive discourse strategies) in Section 5.1.2.

5.1 Interrupted Sentences

5.1.1 Self-Reference

People who talked more about themselves and their group than about the subject matter were more likely to be interrupted. In the given incentive structure of institutional discourse, any pronounced reference to oneself was regarded as an invitation for interruption, as shown in (4) and (5) below:

- (4) → 8454:a 即使我不承認中華民國憲法或者我…
 ‘Even if I don’t recognize the Constitution of the Republic of China or I…’
 8454:b 我剛剛的講法並沒有任何的意思說我不認同這部憲法，我剛剛講法的重點是對這個問題我拒絕表態。
 ‘What I said just now did not mean that I did not agree with the constitution. The point of what I said just now is that I refuse to express my position on this issue.’
- (5) → 6096:a 這樣的體制的確是有點混亂，所以我…
 ‘Such a system is indeed a bit confusing, so I…’
 6096:b 你不覺得林全像小媳婦？是不是？他只是管家而已啊。
 ‘Don’t you think Mr. Lin Quan is like a little daughter-in-law? Is he not? He is just a housekeeper.’

5.1.2 Reasonable Presentation

In a verbal conflict, the side with the better argument is supposed to win; hence, given the incentive structure of the zero-sum game during a political verbal exchange, the attacking side was inclined to prevent the other from clearly presenting his or her argument. Ifs and buts were welcomed weak points ready for exploitation. We observed break points at *ruguo* (如果) ‘if’, *ruguo shuo* (如果說) ‘if’, and *jiaru* (假如) ‘if, in case’, among others, suggesting that arguments introduced with an irrealis were considered weaker because they were less likely to be true or relevant, as shown in (6) and (7) below:

- (6) → 462:a 總統曾經說過不排除任何的可能性，不過如果…
‘The president once said that no possibility is ruled out, but if...’
462:b 可能性高不高？
‘Is the probability high?’

Counterfactuals belonged to the realm of hypotheticals and were often introduced with *fouze* (否則) ‘otherwise’, *buran* (不然) ‘if not’, *buguan* (不管) ‘no matter what’, *jiusuan* (就算) ‘even if’, *chufei* (除非) ‘unless’, *faner* (反而) ‘instead’, and others. We observed a tendency also for counterfactuals to appear in interrupted sentences, as shown in (7) below:

- (7) → 7253:a 我的猜測是，因為兩岸關係條例基本上、原則上的前提是不承認大陸的學歷，除非主管機關…
‘My guess is that the basic and principle premise of the regulations on cross-strait relations is that mainland academic qualifications are not recognized unless the competent authority...’
7253:b 沒有啦，我現在要講的就是，你討厭它可以，你不喜歡它也可以，兩岸關係緊張也可以，即使你認為它是共匪、共產黨都可以，可是現在對岸的北京大學、清華大學是不是比我們的台大排名還在前面，這也是事實。
‘No, what I want to say now is, you can hate it, you can dislike it, and cross-strait relations can be tense, even if you think it is a communist bandit or the Communist Party, but aren’t Peking University and Tsinghua University on the other side of the bank ranked ahead of our National Taiwan University? That’s also the truth.’

5.1.3 Over-Limitation and Pseudo-Objectivism

In the given incentive structure of interpellations, the opposing parties sought to exploit each other's weaknesses and mistakes. This led to the discourse participants avoiding any overly subjective, absolute, or general statements. When adopting a defensive discourse strategy, they tried to appear balanced, objective, specific, and restrictive in their use of language. Hence, we observed many lexical items that were used as downtoners, or hedges, to limit a given proposition in terms of time, subject, certainty, relevance, and so on. A commonly observable lexical item with this discourse function was *muqian* (目前) 'currently', which worked as a protective shield against questions about past or future developments of a certain topic. But it also signaled limited knowledge or responsibility of the speaker. This category also included *duiyu* (對於) 'in regard to', *zhiyu* (至於) 'in regard to', *yixie* (一些) 'some', *bufen* (部分) 'partly', *dabufen* (大部分) 'mostly', *youde* (有的) 'some', *zhuyao* (主要) 'most importantly', *yinggai* (應該) 'should, possibly', *keneng* (可能) 'possibly', *chabuduo* (差不多) 'roughly', *huoxu* (或許) 'perhaps', *yingdang* (應當) 'should', *yuanze shang* (原則上) 'in principle', *zhaoli* (照理) 'theoretically', *bujiande* (不見得) 'not necessarily', *jinkuai* (儘快) 'as fast as possible', *jinzaoyao* (儘早) 'as soon as possible', *jinliang* (儘量) 'try to', *benlai* (本來) 'actually', *qishi* (其實) 'actually', *tanbai* (坦白) 'to be honest', and *dagai* (大概) 'roughly speaking'. An example of this defensive discourse strategy is shown in (8) below:

- (8) → 76:a 有就這部分做一些…
 'If so then on this part do some...'
 76:b 本席在這裡還是要向你提出最嚴正的抗議，如果你的民調要上來，如果你不能對美國、對日本說 NO，只一味在立法院蠻幹…
 'Sir, I still have to lodge my most solemn protest here. If your polls are going to come up, if you can't say NO to the United States or Japan, just blindly do it in the Legislative Yuan...'

Related to the category of subject limiters were words that referred to a specific part or set of a category or topic, such as *yixie* 'some', *bufen* 'partly', *dabufen* 'mostly', *youde* 'some', and *zhuyao* 'most importantly', among others, as shown in (9) below:

- (9) → 3912:a 這個部分沒有…
 'Nothing in this regard...'
 3912:b 碰到你們自己的事情，你們就說跟他無關、跟你無關或跟誰無關。
 'When it comes to your own affairs, you say that it has nothing to do with him,

with you, or with anyone.’

Self-limitation refers to an argument that is aimed at providing counter-proof or negative evidence of one’s own statement and that often signals weakness and invites verbal intrusion. This type of argument marker included *buguo* ‘but’, *danshi* ‘but’, *qishi* ‘actually’, and *tanbai* ‘to be honest’, among others, as shown in (10) below:

(10) → 6679:a 這個個案坦白講…

‘Frankly speaking…’

6679:b 不是個案。

‘Not a case.’

The limiters ‘possibility’, ‘expectation’, ‘certainty’, ‘importance’, and ‘ability’ were also present in the defensive discourse. Signaling a lack of knowledge, credibility, or responsibility, limiters were often exploited for interruptions, such as *yinggai* (應該) ‘should’, *keneng* (可能) ‘possibly’, *huoxu* (或許) ‘maybe’, *yingdang* (應當) ‘should’, *yuanzeshang* (原則上) ‘in principle’, *zhaoli* (照理) ‘reasonably’, *bujiande* (不見得) ‘not necessarily’, *jinkuai* (儘快) ‘as fast as possible’, *jinza* (儘早) ‘as soon as possible’, *jinliang* (儘量) ‘trying to’, and *benlai* (本來) ‘actually’, among others, an example of which is shown in (11) below:

(11) → 2430:a 這件事情其實各方都高度關注，所以我們去看了相關資料，事實上是有一些不是很清楚的地方應該要再…

‘In fact, all parties are paying great attention to this matter, so we went to read the relevant information. In fact, there are some unclear points that should be re…’

2430:b 對於投資案，你不要畫條線把它檔在門口，只因為懷疑它以前是強盜或什麼的，但沒有證據嘛，你要拿出證據來啊!

‘For investment cases, you should not draw a line to file it at the door, just because you suspect that it was a robber or something before, but there is no evidence, you have to show evidence!’

If a speaker lacked specific knowledge about a certain subject or procedure, he or she often retreated to general rules to provide a generic inference. In such cases, the speakers often used

yuanzeshang (原則上) ‘in principle’ or *jiben[shang]* (基本[上]) ‘basically’, as shown in (12) below:

- (12) → 1450:a 這個部分，我們尊重農委會的處理，原則上我們會…
 ‘In this part, we respect the handling of the Council of Agriculture. In principle, we will...’
- 1450:b 剛才農委會主委說可以，是不是？
 ‘Just now the chairman of the Council of Agriculture said yes, didn’t he?’

5.1.4 Over-Confidence

In contrast to the discussion above, words indicating too much confidence and certainty were also exploited for interruptions. These are also often called amplifiers, intensifiers, and boosters. In this study, they comprised a small number of words, which were arguably less relevant to why the sentences were interrupted in the political discourse. However, confidence, indeed, played an important role in the defensive discourse strategies. Words indicating confidence and certainty were, in general, helpful in protecting against interruptions. This category included words such as *yiding* (一定) ‘definitely’ and *dangran* (當然) ‘of course’, as shown in (13) below:

- (13) → 3154:a 目前是考量設在臺南或高雄，會由臺南與高雄雙方的首長做最好的規劃，不論規劃為何，原本臺南或高雄已在進行的工作一定不會…
 ‘At present, it is considered that if it is located in Tainan or Kaohsiung, the heads of both Tainan and Kaohsiung will make the best plan. No matter what the plan is, the work that Tainan or Kaohsiung is already doing will definitely not...’
- 3154:b 你來自高雄，應該知道高雄原本就有一個熱帶醫學中心。
 ‘You are from Kaohsiung. You should know that Kaohsiung originally had a tropical medicine center.’

Another example of over-confidence was words that indicated truth, proof, or the absence thereof, including *queshi* (確實) ‘indeed’ and *bukeneng* (不可能) ‘not possible’, see (14) below:

- (14) → 6696:a 法律要與時俱進，法律不可能一成不變，…
 ‘The law must advance with the times, and the law cannot be immutable,...’
- 6696:b 對，但是我們所謂的溯及既往是以對當事人有利為原則，對不對？

‘Yes, but what we call retroactivity is based on the principle of benefiting the parties, right?’

5.1.5 Evaluation

The data suggest that certain lexical items or expressions related to judgments, either objective or subjective, triggered interruptions more often because evaluative words highlighted the speaker’s judgment about a discourse topic through emotional effort (*nuli*), a difference (*butong*), or something regarded as special (*teshu*), among others. Every objective evaluation was (over-)turned, often for rhetorical purpose, into a subjective statement by the opposition in the political discourse in order to initiate a verbal attack. Evaluative words included *nuli* (努力) ‘with effort’, *butong* (不同) ‘different’, *tebie* (特別) ‘special’, *teshu* (特殊) ‘special’, *kunnan* (困難) ‘difficult’, *danxin* (擔心) ‘afraid’, *yange* (嚴格) ‘strict’, *shiji* (實際) ‘in reality’, *mingque* (明確) ‘clearly’, *zunzhong* (尊重) ‘respect’, *xiwang* (希望) ‘hope’, and *dique* (的確) ‘indeed’, an example of which is shown in (15) below:

- (15) → 155.1:a 政府非常的努力…
‘The government has worked very hard...’
155.1:b 院長同意他用這樣的方式來回答嗎?
‘Does the dean agree with him to answer in this way?’

Pointing out something as special also often cued in others for interruptions, using words such as *butong* ‘different’, *tebie* ‘special’, and *teshu* ‘special’, as shown in (16) below:

- (16) → 6507:a 因為每個人在不同的時候有不同的身分，他必須要做符合他身分的事情，
譬如說…
‘Because everyone has a different identity at different times, he must do things
that match his identity, such as...’
6507:b 你可以保證在未來 8 年中，你不會採取制憲的角度，也不會把你學者的身
分帶到這個地方嗎?
‘Can you guarantee that in the next eight years, you will not adopt a
constitutional perspective, nor will you bring your status as a scholar to this
place?’

Value judgments about something being difficult and worrisome—such as *kunnan* (困難) ‘difficult’ and *danxin* (擔心) ‘afraid’—made up a small subgroup of subjective evaluations inviting interruptions, as shown in (17) below:

- (17) → 4271:a 如果許部長認為這中間有困難的話會跟我解釋，我們也不是每個任命都是…
 ‘If Minister Xu thinks there are difficulties in the process, he will explain to me. Not every appointment is…’
- 4271:b 院長，你也不要害你的部屬，張兆順的名單一開始是他想出來的嗎？
 ‘Dean, don’t hurt your subordinates either. Did Zhang Zhaoshun come up with the list at the beginning?’

5.2 Interrupting Sentences

Interrupting sentences were characterized as such because they disrupted a statement of an interactant. In comparing all the interrupting sentences, we observed certain recurrent features and referred to them as “interrupting-keywords.” The interrupting keywords were not limited to specific word classes. Since they appeared anywhere in a sentence, they were not understood as “causing the interruption,” but rather as linguistic patterns that were naturally preferred, or manifest, when a speaker realized the speech act of interruption.

5.2.1 Negation

Arguably, one of the most prominent functions of the interruptions was to disagree with an opponent. This was indicated by negative particles, such as *buyao* (不要) ‘do not’, *bu* (不) ‘no’, and *mei* (沒) ‘no, not’, among others. Direct negation of a verb, other than implicit negation, was the most frequent type, such as *fandui* (反對) ‘oppose’ and *kunnan* (困難) ‘difficult’. There were 297 instances of direct verb negation in the interrupting sentences, roughly one-third more than in the interrupted sentences (207). The words *buyao* (不要) ‘do not’, *buhui* (不會) ‘will not/cannot’, *buneng* (不能) ‘cannot’ and *buxing* (不行) ‘cannot’ also showed this tendency, but with a decreased contrast, which referred to opposite frequency shifts (increased vs. decreased) between the interrupted and the interrupting sentences. As such, negation was closely related to adversatives and opposition. Examples of this offensive discourse strategy are shown in (18) and (19) below:

- (18) 1499:a 我們來評估，那個部分…
‘Let’s evaluate, that part…’
→ 1499:b 你們不要評估了，你們已經評估很久了。
‘You don’t need to evaluate it, you have been evaluating it for a long time.’
- (19) 4304:a 我們不能有…
‘We can’t have…’
→ 4304:b 難道你們不會適度告訴他們嗎？
‘Wouldn’t you tell them in moderation?’

5.2.2 Adversatives and Opposition

Adversative words imply rejection, protest, or contrast of opinion, and in the context of the interpellations, they signaled the interrupter’s opposition. Words that fell into this group included *que* (卻) ‘yet, but’, *keshi* (可是) ‘but’, and *zhi* (只) ‘only’. In many cases, the burden of signaling opposition to something did not fall on these words alone but also relied on a “combined occurrence,” in which the discourse function of an utterance (i.e., showing opposition) was fulfilled by a set of words that all occurred together in the same sentence. This category included *que* (卻) ‘yet, but’, *keshi* (可是) ‘but’, *zhi* (只) ‘only’, *xin* (新) ‘new’, *jiu* (就) ‘(grammatical particle)’, and *zhiyao* (只要) ‘if only’, an example of which is shown in (20) below:

- (20) 8034:a 當然稅法和憲法的關係就是憲法第十九條，人民有依法律納稅之義務…
‘Of course, the relationship between the tax law and the constitution is Article 19 of the constitution, and the people have the obligation to pay taxes in accordance with the law…’
→ 8034:b 是，可是納稅的主體其實不是人民啊！
‘Yes, but the taxpayers are not actually the people!’

5.2.3 Superlatives

Interruptions are considered a rhetorical device that indicates strong emotion, opposition, or involvement. Superlatives naturally support the sense of contrast and opposition. In this context, they have been discussed under the label “extreme case formulation” by Pomerantz (1986). We

observed many instances of superlatives in the interrupting sentences, using words such as *lian* (連) ‘even’, *zui* (最) ‘most’, *tai* (太) ‘too, most’, and *juedui* (絕對) ‘absolute(ly)’, among others. Examples of this offensive discourse strategy are shown in (21) and (22) below:

- (21) 8491:a 我的印象好像沒有，最多大概是類似用行政命令去…
 ‘I don’t seem to have any impressions, at most it is probably similar to using administrative orders...’
 → 8491:b 連行政命令都沒有!
 ‘There are absolutely no administrative orders!’
- (22) 6986:a 這在台灣的話，…
 ‘If this is in Taiwan,...’
 → 6986:b 這個話題談太久了，其實我不是在質疑你的學術地位或法學素養。
 ‘This topic has been talked for too long. Actually, I am not questioning your academic status or legal literacy.’

5.2.4 Questions

Out of the 1,089 interruption pairs, sentences that ended in a question mark were the most common category of the interrupting sentences (456 instances, 42%), closely followed by statements (442 instances, 41%). Interpellations are all about asking questions. Both real and rhetorical questions are powerful rhetorical devices. In the institutional discourse, questions were often used to perform the discourse functions of showing disrespect, power, and aggressive verbal attacks. This category included words such as *ma* (嗎) ‘MA-particle’, *ne* (呢) ‘NE-particle’, *weishenme* (為什麼) ‘why’, *weihe* (為何) ‘for what’, *na* (哪) ‘which’, *shenme* (什麼) ‘what’, *duoshao* (多少) ‘how much/many’, *haobuhao* (好不好) ‘all right?’, *nengbuneng* (能不能) ‘can (you)?’, and *huibuhui* (會不會) ‘would (you)?’. Yes/No questions ending in *ma* (嗎) were five times more common than questions ending in *ne* (呢), and they were responsible for a quarter of all the questions. An example of this offensive discourse strategy is shown in (24) below:

- (24) 2720:a 長照的部分因為跟原民會的部分有相關…
 ‘Because the long-term care policy is related to the Council of Indigenous Peoples...’

- 2720:b 我們今天沒有特別邀請，以後我們再弄一個專案報告好嗎?
'We have no special invitation today. We do another project report in the future, alright?'

Questions formed with *haobuhao*, *nengbuneng*, and *huibuhui* were more often related to disrespect and showing power, as shown in (25) below:

- (25) 1671:a 我會誤會呂委員是因人設事來改制度，這是不好的，一個制度要去改是因為...
'I would misunderstand that Commissioner Lu changed the system because of the establishment of personnel. This is not good. A system needs to be changed because...'
- 1671:b 我是在野黨的，你如果不敢說，怕影響黨內的和諧，我幫你提，好不好?
'I'm from the opposition party, if you dare not say it, for fear of affecting the harmony within the party, I'll help you mention it, okay?'

5.2.5 Direct Address

In the interrupting sentences, we often observed certain personal pronouns used to directly address the opposing party. More often, the interrupter called the addressee by his or her professional title (i.e., *buzhang* (部長) 'minister' and *yuanzhang* (院長) 'dean'). This category also included words such as *ni* (你們) 'you', *nimen* (你們) 'you' (pl.), and *zhuwei* (主委) 'chairman'. Examples of this offensive strategy are shown in (26) and (27) below:

- (26) 7375:a 對於這個問題，其實我們的社會已經討論很多了，個人覺得我的看法如何其實已經不是那麼的重要，我個人的選擇...
'In fact, our society has already discussed this issue a lot. I personally feel that my opinion is not so important anymore. My personal choice...'
- 7375:b 你個人的選擇是什麼?
'What are your personal choices?'
- (27) 3907:a 以後我們不要發生這種情況，這是沒有錯，但是...
'Let's make sure that it won't happen again in the future, it is not wrong, but...'
- 3907:b 院長，人要誠實，我所認識的林全，不是像你這樣在耍嘴皮子的。

‘Dean, people have to be honest. The Lin Quan I know is not playing tricks like you.’

6. Summary and Conclusion

Institutional discourse differs from other types of communication in that its incentive structure is clearly defined as confrontational, and it rewards aggressive linguistic behavior, which is categorized by forms of defensive and offensive discourse strategies and is associated with certain linguistic patterns at both the parts-of-speech level and the semantic-patterns level. As the corpus-based analysis has shown, both levels reflected strategies of interruptions. In contrast to interruptions in other settings, which can be explained by cues and speech markers, the interruptions during the political interpellations in the current study were not invited or semi-planned. Rather, the interruptions happened in an incentive structure that rewarded the exploitation of the opponent’s weaknesses in his or her argument or presentation. Expressions related to self-reference, a reasonable presentation of an argument, pseudo-objectivism, displays of confidence, or any word that could be interpreted as having a subjective viewpoint are common categories of interrupted sentences. From the perspective of the opponent, these categories represent weakness and invite interrupting attacks. Each of these construed weak points has been statistically associated with increased or decreased frequency shifts in the keywords and semantically with discourse functions.

In the interruptions, pronouns, conjunctions, and adverbs were overrepresented, given their numbers in the overall corpus. Within the entire corpus, pronouns made up less than 1%, but they represented 10% of all the statistically significant keywords of interruption. In terms of conjunctions, they made up around 1% and represented 9% of the keywords. For adverbs, they made up 5% and represented 22% of the keywords. Conversely, nouns, verbs, and adjectives were underrepresented. Nouns comprised the largest group in the entire corpus (51%) but contributed only about 29% to the keywords. Verbs accounted for 40% in the corpus but only 18% of the keywords were verbs. Adjectives played no role at all (about 1%)—not a single adjective was a keyword.

Statistics can explain the frequency effects only to a certain degree. The more interesting question is, why were discourse function words such as conjunctions not equally distributed across sentence types, but in fact, showed significant differences? Moreover, to what degree were conjunctions, pronouns, and adverbs, as well as some nouns and verbs, related to discourse strategies? These questions required a second tier of analysis.

In this second step, we used the incentive structure in order to explain the shifts in frequencies. We differentiated between offensive discourse strategies, which tended to be more

associated with interrupting sentences, and defensive discourse strategies, which were more associated with interrupted sentences.

Important words discussed in this study were the keywords in the interrupted sentences, which were often related to downtoners and expressed pseudo-objectivity, such as *muqian* (目前) ‘currently’, *qishi* (其實) ‘actually’, *keneng* (可能) ‘possibly’, and *bijiao* (比較) ‘comparatively’; to over-confidence, such as *dangran* (當然) ‘of course’ and *yiding* (一定) ‘definitely’; to nouns, such as *shishi* (事實) ‘fact’; to verbs that mainly had discourse functions, such as *jiang* (講) ‘talk’; to express subjective evaluation, such as *xiwang* (希望) ‘hope’; and to self-reference, making excessive use of first-person pronouns, such as *wo* (我) ‘I, me’ and *women* (我們) ‘we, us’. Conjunctions also appeared significantly more often in the interrupted sentences, especially when introducing subclauses that indicated reason or counterfactuals, such as *yinwei* (因為) ‘because’, *suoyi* (所以) ‘therefore’, *ruguo* (如果) ‘if’, and *danshi* (但是) ‘but’.

In the interrupting sentences, on the other hand, the higher-frequency keywords were second-person pronouns that were used to directly attack an opponent, such as *ni* (你) ‘you (sg.)’ and *nimen* (你們) ‘you (pl.)’, and adverbs related to counter-attack pseudo-objectification, such as *xinzai* (現在) ‘now’, in opposition to *muqian* (目前) ‘currently’. Words that indicated opposition or negation were particularly prevalent in the interrupting sentences, such as *buyao* (不要) ‘do not’, *bu* (不) ‘not’, and *meiyou* (沒有) ‘do not have’. These words were also related to speech acts in general, such as *wen* (問) ‘ask’ and *huida* (回答) ‘answer’, and to words that indicated a conclusion, mostly *suoyi* (所以) ‘therefore’. These examples demonstrate that the keywords were organized by semantic fields and were related to discourse functions.

Taken together, interruptions in institutional discourse can be explained, at least partially, by frequency patterns and semantic patterns embedded in a competitive incentive structure. Interruptions are a multi-layered phenomenon that works at different levels simultaneously, as illustrated in Figure 5 below:

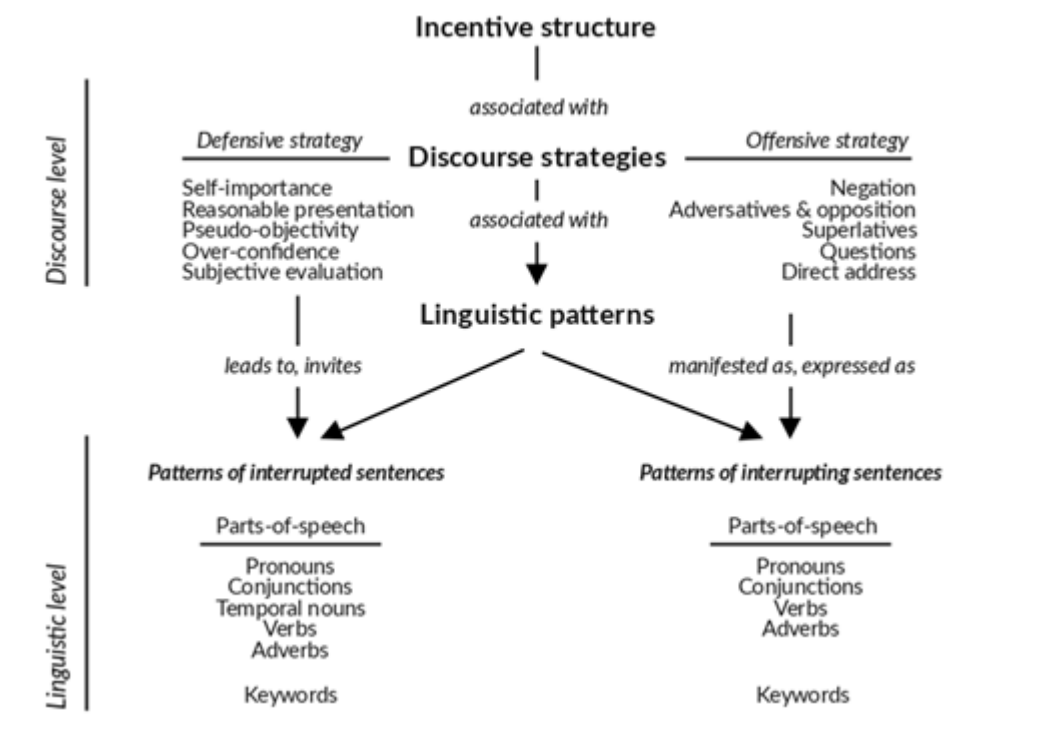


Figure 5. Discourse and linguistic levels of interruption

This paper mainly focused on depicting the keywords involved in interruptions during parliamentary discourse. However, interruptions are a complex linguistic phenomenon. Other underlying mechanisms, such as the effect of different stances of the interlocutors, the intentions of the interlocutors, and even the existence of interruptive constructions, are also intriguing topics. Indeed, they are beyond the scope of this paper, so we will leave those topics for future studies.

Acknowledgments

We would like to express our sincere gratitude to the two anonymous reviewers for their constructive comments that helped improve the quality of the previous manuscript. We also would like to thank Andrew H.C. Chuang for the first proofreading and Prof. Siaw-Fong Chung for her inspiring inputs and support for this publication.

References

- Bazzanella, C., Caffi, C., & Sbisà, M. (1991). Scalar dimension of illocutionary force. In I. Z. Žagar (Ed.), *Speech acts: Fiction or reality* (pp. 63-76). IPrA Distribution Centre for Yugoslavia Institute for Social Sciences.

- Duncan, S. D. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2), 283-292. <https://doi.org/10.1037/h0033031>
- Ferguson, N. (1977). Simultaneous speech, interruptions and dominance. *British Journal of Social and Clinical Psychology*, 16(4), 295-302. <https://doi.org/10.1111/j.2044-8260.1977.tb00235.x>
- Fraser, B. (1990). An approach to discourse markers. *Journal of Pragmatics*, 14(3), 383-398. [https://doi.org/10.1016/0378-2166\(90\)90096-V](https://doi.org/10.1016/0378-2166(90)90096-V)
- Goldberg, J. A. (1990). Interrupting the discourse on interruptions : An analysis in terms of relationally neutral, power- and rapport-oriented acts. *Journal of Pragmatics*, 14(6), 883-905. [https://doi.org/10.1016/0378-2166\(90\)90045-F](https://doi.org/10.1016/0378-2166(90)90045-F)
- Hutchby, I. (1996). *Confrontation talk : Arguments, Asymmetries and Power on Talk Radio*. Lawrence Erlbaum.
- Jefferson, G. (1978). Sequential aspects of storytelling in conversation. In J. Schenkein (Eds.), *Studies in the organization of conversational interaction* (pp. 219-248). Academic Press.
- Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.
- Ma, W. Y., & Chen, K. J. (2003). Introduction to CKIP Chinese word segmentation system for the First International Chinese Word Segmentation Bakeoff. In *Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing*, 168-171. <https://doi.org/10.3115/1119250.1119276>
- Nor, S. (2012). Discourse markers in turn-initial positions in interruptive speech in a Malaysian radio discourse. *Multilingua*, 31(1), 113-133. <https://doi.org/10.1515/mult.2012.005>
- Oreström, B. (1983). *Turn-taking in English conversation*. Gleerup.
- Pomerantz, A. (1986). Extreme case formulations: A way of legitimizing claims. *Human Studies*, 9(2), 219-229. <https://doi.org/10.1007/BF00148128>
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696-735. <https://doi.org/10.2307/412243>
- Schegloff, E. A. (2002). Accounts of conduct in interaction: Interruption, overlap and turn-taking. In J. H. Turner (Ed.), *Handbook of sociological theory* (pp. 287-321). Plenum.
- Signes, C. (2000). A genre based approach to daytime talk on television. *SELL monographs 1*. Universitat de València.
- Stainton, C. (1987). Interruptions: A marker of social distance? *Nottingham: Occasional Papers in Systemic Linguistics*, 2, 75-135.
- Tannen, D. (1984). *Coherence in spoken and written discourse* (Vol. 12). Praeger.
- Waltereit, R. (2002). Imperatives, interruption in conversation, and the rise of discourse markers: A study in Italian guarda. *Linguistics*, 40(5), 987-1010. <https://doi.org/10.1515/ling.2002.041> Available at: https://www.researchgate.net/publication/228856522_Imperatives_interruption_in_conv

ersation_and_the_rise_of_discourse_markers_A_study_of_Italian_guards/link/0a85e53317fff5ca11000000/download

Wiemann, J., & Knapp, M. (1975). Turn-taking in conversations. *Journal of Communication*, 25(2), 75-92. <https://doi.org/10.1111/j.1460-2466.1975.tb00582.x>

Wilcoxon Signed-Ranks Test Calculator. Available at:

<https://www.socscistatistics.com/tests/signedranks/default.aspx> (Accessed January 24, 2022).

Appendix A. Supplementary material

The keyword list of this study is available via the following link:
<https://www.space.ntu.edu.tw/navigate/s/462D30BB20BF4E8B826F0610FE6C3052QQY>

以社群媒體語言建構深度學習模型：

以「校正回歸」為例

Constructing a Deep Learning Model Using Language in Social Media: The Case Study of ‘Retrospective Adjustment’

段人鳳*、邱淑怡+、劉慧雯#

Ren-feng Duann, Shu-I Chiu, and Hui-Wen Liu

摘要

本研究以台灣新冠肺炎期間首度出現的「校正回歸」一詞相關的臉書貼文為語料，進行人工情感分析與模型預測。我們對 6,917 筆語料進行人工標記，並將這些標記完成的語料分成 70% 和 30%，以 BERT-Chinese 之預訓練模型(pre-trained model)利用 70%的語料進行微調機制(fine tune)，再以微調後的模型預測剩餘的 30%語料，並加以比對人工標記和模型預測的結果，試圖從語言特徵找出兩者間差異的可能原因。研究結果顯示，在人工標註為中立的貼文中，模型有較好的預測能力，正確率達 0.81；而人工標註為正向和負向的貼文中，模型的預測能力較差，分別為 0.64 和 0.63。進一步觀察人工標註和模型預測的差異，人工標註為負向而模型預測為正向的有 0.23，乃所有錯誤之最，其次為

* 國立臺東大學通識教育中心

Center for General Education, National Taitung University.

E-mail: rduann@nttu.edu.tw

+ 國立政治大學資訊科學系（通訊作者）

Department of Computer Science, National Chengchi University. No. 64, Sec. 2, Zhina Road, Wenshan District, Taipei City, Taiwan 116011, R.O.C.

Tel: 886-(0)2-29393091 ext. 88112. Fax: 886-(0)2-22341494

E-mail: sichiu@nccu.edu.tw

國立政治大學新聞系

Department of Journalism, National Chengchi University.

E-mail: huiwen@nccu.edu.tw

人工標註為正向而模型預測為中立的貼文，0.22。我們逐筆檢視這兩大類貼文，歸納出 7 類負向情感的語言特徵及 4 類正向情感的語言特徵。在檢視語言特徵時，研究者亦發現，由於本文所搜集之語料具有高度的公共性與政治性，僅討論貼文內容有時不易判斷意義，還需考慮貼文者身份，此亦可能影響了機器預測的正確率。我們主張，社群媒體的語言有別於當下模型訓練使用的資料集，且貼文者常常使用表情符號或標點符號來表達情感，未來我們將發展適合台灣的社群媒體語意的預測模型，以期提升模型預測的正確率。

Abstract

This research, which used Facebook posts related to the term “retrospective adjustment” in Taiwan as the corpus, manually coded the sentiments of 6,917 posts. Randomly dividing the dataset into two subsets for training (70%) and testing (30%) and using the Chinese pre-trained BERT model as the foundation, we trained and fine-tuned the model with the training dataset and ran the fine-tuned model to predict the sentiments in the test dataset. We then compared the results of the manual coding and model prediction to explain the differences from the perspective of linguistic features. The results indicated that the model performed better for the posts manually coded as “neutral,” with an accuracy of 0.81, while the accuracies of model prediction were only 0.64 and 0.63 for the posts manually coded as “positive” and “negative,” respectively. Regarding inaccuracy, the posts manually coded as “negative” but predicted by the model as “positive” and those manually coded as “positive” but predicted by the model as “neutral” ranked the highest (0.23) and the second highest (0.22), respectively. Examining the linguistic features of the two groups of posts, we identified seven categories of linguistic features that, we claim, led to “negative” coding and four categories that led to “positive” coding. Moreover, both groups contained posts that could not be coded accurately without knowledge of the news and the Facebook account owners’ political/social inclinations, which was attributed to the posts’ high relatedness to the general public and the politics of Taiwan. Considering that the language used in social media is different from the language employed to train current models, and that Facebook users frequently use punctuation marks and emoticons to express their moods, we argue that there is a need to develop a model for social media.

關鍵詞：社群媒體、深度學習、校正回歸、情感分析、自然語言處理

Keywords: Social Media, Deep Learning, Retrospective Adjustment, Sentiment Analysis, Natural Language Processing

1. 緒論 (Introduction)

社群媒體(social media)的發展,自 2006 年起,進入一個全新的樣貌。包括 Plurk、Facebook, 以及 Twitter 在內,透過發展出能夠由使用者自行建立的「朋友清單」(friends' list),以及更為簡便的貼文介面,人們透過社群媒體發表意見,成為 21 世紀重要的媒體活動。2007 年,第一代 iPhone 手機問世,1992 年即已在商展上展出的智慧型手機從此進入商轉。隨著網路系統建置覆蓋率的提升,行動通訊徹底改變媒介使用的時間性,人與人之間的意見交換更為迅速。同時,也因為貼文門檻的大幅降低,使得社群媒體上刊載了大量由一般人表達的意見。Keen (2007)認為這些由「業餘者教派」產製並分享的意見會扼殺我們的文化價值,主張不該放任「任何人」對任何主題隨意發表評論。

然而,不論論者是否同意,網際網路基礎架構,加上降到極低的參與門檻,再加上行動通訊載具三方的共同發展,促成的全民貼文、全民分享行為,已經是不爭的事實。光是社群媒體龍頭臉書(Facebook)在 2021 年 1 月底發佈的報告就指出,即使在劍橋分析(Cambridge Analytic)公司爆發個資外洩事件之後,臉書的社會聲望下挫,但 2020 年 12 月底,每日仍有 18.4 億個活躍帳號;尤其,最愛使用臉書的地區,分別是拉丁美洲與東南亞(Facebook, 2021)。

在這個大規模的媒體使用活動中,產生了大量「使用者產生內容(user-generated content, UGC)」。這些內容透過朋友清單迅速地將個別意見傳送給特定社群網絡,形成類公共論壇(public sphere)。是以,自從 2010 年之後,掌握並分析社群媒體上的材料,就成為了解一般人意見、態度,乃至於輿論風向的關鍵。由於 UGC 的生產快速而量大,因此,研究者必須透過資訊工具,才有可能迅速掌握內容,同時也才能進行資料清洗與分析等工作。其中,最常見的就是語料分析方法。

本研究以新冠肺炎期間,環繞著指揮中心於 2021 年 5 月 22 日開始使用的「校正回歸」一詞所產製的臉書貼文為語料,先進行人工情感分析,再將完成人工標記之語料切分為 70%和 30%,前者為訓練資料集(training dataset),後者為測試資料集(test dataset)。以前者訓練 BERT-Chinese 模型,再以該預訓練模型預測後者的情感,並比對人工標註和模型預測的結果,以期從人工標註和模型預測的差異找出不同情感向度(負向和正向)的語言特徵,以為後續模型發展提出我們的建議。

本研究主要回答下列兩個問題：

- 一、人工標註和模型預測之差異為何？
- 二、負向／正向情感的語言特徵為何？

本文第 2 章回顧自然語言處理與社群媒體語言特徵等相關研究；第 3 章說明本研究採用的語料、標記方法與使用的模型；第 4 章陳述語料分析結果與討論,第 5 章則說明本研究的結論、限制與未來方向。

2. 文獻回顧 (Literature Review)

2.1 自然語言處理 (Natural Language Processing, NLP)

自然語言處理是人工智慧(artificial intelligence)與語言學的結合領域，此領域探討如何處理及運用自然語言。自然語言處理分為認知、理解、生成三個步驟：認知和理解是讓電腦把輸入的語言變成有意義的符號與關係，再依據目的進行處理，最後利用演算法生成最後的結果。近年來電腦硬體設備效能不斷地提升，2014年由英國 Google DeepMind 開發的 AlphaGo 人工智慧圍棋軟體，它由 Silver *et al.* (2016) 訓練，為第一個無需讓子之下擊敗圍棋職業九段棋士的電腦圍棋程式。AlphaGo 採用兩個深度學習(deep learning)的類神經網路(neural network)及搜尋機制來選擇落子，並進一步使用強化學習(reinforcement learning)加以改善。在這股風潮引領之下，深度學習再度成為熱門話題，透過類神經網路的多層架構的學習機制，讓模型更接近人類大腦可以自發學習進行訓練。這樣的技術也應用於文字的語意分析(semantic analysis)，運用大量閱讀文本資料後，可進行語意分析將文本資料區分類型、對文本資料進行摘錄、或進行文本內容的預測，例如：讓電腦閱讀大量的金庸武俠小說後，模型便可寫出另一本武俠小說。

然而並非每位研究者都能取得大量文本進行模型建構，因此，Devlin *et al.* (2018) 提出自然語言處理的預訓練的技術，由 Google 於 2018 年發布的基於變換器的雙向編碼器表示技術(bidirectional encoder representations from transformers, BERT)，這是一種預訓練語言的方法，Devlin *et al.* 以大量文本語料庫（如：維基百科）訓練一個通用的語言理解的模型。BERT 的模型一開始以英文文本進行預訓練，之後由 Cui *et al.* (2019) 等人發展出以中文訓練的 BERT-Chinese，他們採用大量的中文文本建立模型，提出的模型能處理 NLP 的中文句子和文本。本研究將利用 BERT-Chinese 預訓練的模型，再以我們的資料訓練模型，讓預訓練模型進行微調(fine-tuning)機制，建構分析臉書貼文情感之模型。

情感分析向來受到自然語言處理學者關注，社群媒體的貼文提供了這些學者豐富的語料。Roberts *et al.* (2012) 探討 Twitter 上推文的情感分類，他們蒐集了 14 個主題的推文，如情人節、2010 世界盃、2012 美國總統大選、琳賽蘿涵...等，將憤怒、厭惡、恐懼、喜悅、愛、悲傷和驚訝等七種情緒，分別歸類為正向、負向及未預期(unexpected)三大類型¹。Nissim & Patti (2017) 認為語意層面至關重要，他們回顧了自然語言學界發展之情感分析的語意資源：除了帶有正／負向標記或情緒標記的詞典，也包含在概念層次的情感／情緒分析上，呈現與多字詞語(multiword expressions)相關的語意、概念和情感訊息。他們認為，單字層次的分析顯然已不足以判定發言人的情感，因此出現將分布語意學(distributional semantics)－在相同前後文出現的字詞應有類似的意義－作為情感分析的判斷依據。Tang *et al.* (2014) 提出神經網路來學習能將正／負情感編入字串(n-gram)的特定情感詞嵌入(sentiment-specific word embedding, SSWE)，並以推特貼文訓練機器。以屬

¹ Roberts *et al.* (2012) 的正向類別為喜悅和愛，負向類別為憤怒、厭惡、悲傷和恐懼，未預期類別則包含恐懼與驚訝。

性為基礎的情感分析(aspect-based sentiment analysis)需要辨識特定實體，以及實體與事件之間的關聯，Socher *et al.* (2013)推出了建置於文法結構之上的回歸神經網路模型(Recursive Neural Tensor Network)，並以包含 20 多萬短語的細緻情感標籤的情感結構樹資料庫(Sentiment Treebank)來訓練這個模型，Socher *et al.*主張這個模型較以往方法能更正確地判斷單句的正／負情感，並能更正確預測短語的細緻情感標籤。

在探討 COVID-19 和社群媒體的互動方面，Wang *et al.* (2020) 以新浪－微博(Sina Weibo)上與 COVID-19 相關貼文為資料集，採用無監督 BERT 模型對情感類別（正向、中立和負向）進行分類，並使用「詞頻－逆文檔頻率(term frequency-inverse document frequency, TF-IDF)」模型匯總貼文的主題，再進行趨勢分析和主題分析，以識別負向情緒的特徵。Lu *et al.* (2021) 也依據微博上 COVID-19 主題的貼文進行情感分類，有別於 Wang *et al.* (2020)，Lu *et al.* 將情感分為對某特定對象的「批判」與「支持」兩類。利用 BERT 模型微調的技術，他們探討在社群媒體上公眾情緒(public sentiment)如何隨著 COVID-19 的傳播而演變，進而預測貼文的情緒分類。Singh *et al.* (2021)則探討 COVID-19 對社會生活影響的情緒分析，因疫情而出現的社交距離，促使人們迅速地轉向在社群媒體上表達意見與渴望，該論文以 BERT 模型對 Twitter 上的推文進行情感分析，並將情感分類為正向、負向和中立，以探究人們的心理狀態。

Jain *et al.* (2022)以 BERT 模型為基礎，加入卷積神經網路(convolutional neural networks, CNNs)，提出新的模型「BERT 擴張卷積神經網路(BERT dilated convolutional neural networks, BERT-DCNN)」，該模型包含情感知識庫，建構概念性層級(concept-level)的情感分析，他們分析美國國內及國際航空公司的消費者評論，評論內容亦分為正向、負向及中立三種類別。

上述文獻顯示，無論是社群媒體的推文或是消費者評論，多數研究採用正向、負向及中立作為語意分析的分類，本研究亦將使用正向、負向及中立作為情感類別。

2.2 社群媒體的語言特徵 (Linguistic Features on Social Media)

基於使用者產生內容的特性，臉書上的互動吸引了語言學－特別是言談分析－學者的目光。Tannen (2013)探討電子媒體上的語言使用，她主張媒體的選擇本身便傳達了「後設訊息」，即「『言談的主題即對話者之間的關係(‘the subject of discourse is the relationship between the speakers’)』…說話者如何傳達訊息，而聽話者如何解讀訊息」(p. 101)。Tannen 主張包含臉書、即時通訊等新媒體上的互動與口語對話非常雷同，使用者之大寫字母、重複驚嘆號或問號、重複單字或母音字母等「熱忱標記(enthusiasm marker)」，類似口語對話中的提高語調、增強情感，以及拉長音節等情緒表達；使用者轉貼連結或簡短回覆，類似口語表達中說話者不直接傳達的訊息；回應速度則類似口語交談中的對話步調與停頓…等，這些特徵皆凸顯新媒體上的溝通與口語對話極為相似。Maíz-Arévalo (2015)研究了英文與西班牙文臉書上朋友間的嘲弄(jocular mockery)，發現兩國的臉書使用者最常使用拼字變化（如表情符號、重複字母、重複驚嘆號或問號、模擬他國語言腔調等）來嘲弄貼文作者，也透過「公式化詞彙（如驚嘆詞、習慣用語等）」、反諷、轉移話題，誇

張詞彙等來傳達嘲弄。Taber (2016)探討喀麥隆境內英語使用者在臉書、即時通訊等社群媒體上的幽默，她歸納出 11 類幽默的表達：(1) 違逆 Grice 準則、(2) 拼字變異（如重複字母或單字、玩弄同音字等）、(3) 音效（如狀聲詞、頭韻或尾韻、共鳴等）、(4) 重複標點符號、(5) 幽默軼事（如新聞或故事）、(6) 笑聲、(7) 玩弄身分（將傳統命名轉移到線上互動、調整真實姓名使之看起來新潮，或縮寫姓名）、(8) 表情符號（包含現成的表情符號和特殊標點符號或字母）、(9) 誇飾、(10) 隱喻和(11) 其他（如無規則可循的符號組合）等。Ye (2019)則聚焦中文社群媒體上，「被」字句的創新用法及其傳達的諷刺意味，透過語言、文化和社會分析，作者指出：傳統被字句可不提及施事者，及其凸顯受事者受負面影響之特質，讓使用者能嘲弄未被指名的施事者／權威者，並表達使用者對施事者／權威者的負面態度。

學者也採批判言談分析的角度探討臉書互動如何形塑、維持、強化內部團體(in-group)和對抗外部團體(out-group)。Morin & Flynn (2014)探討美國的新興政黨—茶黨—支持者如何在 2010 年「催票周末(“Get Out The Vote” weekend)」期間運用臉書來建構、維持身分認同。他們研究了三位雖隸屬於共和黨、但強調茶黨身分的參議員候選人的官方臉書頁，並且發現支持者的回饋可分為兩大主題：攻擊與鼓勵，即攻擊威脅，並鼓勵與候選人團結一致。攻擊的語言包含負面評價用語、貼標籤（或譯為「咒罵法」）、誇飾法等，這些都是藉由攻擊外部團體來堅定內部團體的身分認同；鼓勵的語言包含支持者使用準社會接觸(para-social contact)與候選人互動，例如在臉書上直接與候選人對話，使用 Tannan 提出的熱忱標記，否認選情落後，為候選人辯護等。Al-Tahmazi (2015)探討伊拉克臉書上對政治人物與事件的（再）定位如何使政治討論兩極化。透過分析評論者（即臉書使用者）在臉書上如何將政治人物與其作為正當化／去正當化，評論者將政治討論從議題導向轉變為人物導向，同時，評論者也在與其他評論者的互動中，建構了自己的社會—政治認同，並根據言論，將自己與其他評論者放入對立的線上社群，即內部團體和外部團體；換句話說，臉書使用者之間的線上互動為實際生活的延伸，他們臉書上的政治認同和他們實際生活上的種族、教派或文化群體的屬性交互影響，形成互為敵對的政治陣線。Chibuwe & Ureke (2016)聚焦辛巴威境內兩敵對陣營在 2013 年的選舉中如何攻擊彼此，探討臉書自由來去且匿名的特性如何讓兩敵對陣營攻擊彼此，且導致兩極化，他們主張，即使臉書提供了另類公領域，這個公領域也可能被缺乏理性的辯論粉碎。

3. 研究方法 (Methodology)

3.1 資料搜集與標記 (Data Collection and Coding)

為了在一定範圍內針對社群媒體使用者的語言使用型態進行較為細緻的討論，同時考慮解釋本身的有效性以及可概推能力，本研究選擇臉書為搜集語料之平台。我們利用臉書官方資料蒐集器 CrowdTangle，擷取 2021 年 5 月 22-25 日（共計 4 日）公開資訊中使用或提及「校正回歸」字樣的貼文（包括：粉絲專頁、公開社團與個人帳號的公開貼文）共計 6,917 筆，交由三位標記者標註「正向」(Positive, 以下簡稱 P)、「負向」(Negative,

以下簡稱 N)，以及「中立」(Neutral, 以下簡稱 Neu)三種記號之一。這三位標記者中，第一、二位為傳播背景，第三位為語言學背景。

本文所稱標記，是針對 CrowdTangle 擷取檔案中，登記為 text/message 的欄位進行判讀。我們的標記分為兩階段，第一階段為三位標記者分開標記，我們參考新聞學的「5W1H」(who, when, where, what, why, how)表示事實資訊的主要面向，若貼文僅呈現 5W1H，我們便標記為中立²；我們亦參考陳韋帆、古倫維(2018)發展的中文情感語意分析套件，判斷詞彙的情感向度，若貼文表達正向情感者（如：讚揚、感謝、希望、改善、積極作為…等），我們標註為正向，反之則標註為負向（如：指責、驚嚇、失望、恐懼、擔憂…等）。第二階段，則由第三位標記者審視三位標記者的結果，並給予貼文最終判斷。我們的情感標記採多數決，當兩位或三位標記結果相同時，即確定該貼文的標註結果。

本研究以「校正回歸」為個案，除了因為研究進行期間正值 COVID-19 疫情在台灣首度爆發大規模本土感染，中央疫情指揮中心(CDC)宣佈全國進入三級警戒狀態，相當程度限制外出的情況下，正是社會人心惶惶，一般人對疫情資訊有著高度需求的時刻之外，我們也觀察到「校正回歸」一詞在 5 月 22 日 CDC 例行記者會上首度從衛福部長陳時中嘴裡說出時，舉國嘩然。當天之後，不論是電視上的政論性節目、社群媒體上的意見領袖(key opinion leader)，或者一般庶民，都出現了許多對此詞彙的闡述、評論與解讀。這顯示，這個詞彙雖有學理上的定義，但在市民社會中，卻呈現多樣多元的闡述意義。不僅如此，由於臺灣社會剛經歷 2018 年底九合一大選、2020 年初總統大選、2021 年公民投票等事件，整個社會處於高度社會政治化的狀態中。這個狀態使得「疫情治理」這個公共衛生事件，也成為政治攻防的關鍵場域。由此，我們可以預見中央疫情指揮中心的資訊，也容易政治化。在此時對「校正回歸」這個關係染疫人數的詞彙進行討論，處理的也就不只是一個統計詞彙，更是一個雜揉了政治與社會輿論氣候的詞彙。藉由對這個詞彙的資料搜集、標記與自然語言處理，我們可以透過在社會生活中呈現多義狀態的詞彙，一窺人工標記／解讀，與自然語言處理「解讀」同一批測試資料集的異同，並進一步提出對社群媒體使用者語言實踐和（繁體中文）自然語言處理的調校方式。

3.2 自然語言處理模型 (The NLP Model)

本研究基於自然語言處理方式，採用遷移學習的技術，以 BERT-Chinese 作為模型架構，該模型利用 Wikipedia 上（包含繁體中文及簡體中文）的文章，有 2.1 萬個詞彙，參數數量約 97M (Cui *et al.*, 2019)。本研究藉由訓練和微調技術建構屬於臉書議題的語意模型，

² 在新聞學上，以「5W1H」(who, when, where, what, why, how)表示事實資訊的主要面向，在純淨新聞寫作中，聚焦這六個面向，可以簡短但精準傳達新聞事件的核心內容，提供讀者必要的資訊。因此，5W1H 乃是新聞必要內容。本文將僅涉及 5W1H 的貼文視為中立，因為這表示貼文有、且僅有傳遞資訊的能力。同時，既有的文獻顯示，若欲發展機器下標、撰寫導言等智能模型，也經常以 5W1H 做為模仿的架構（參見：鍾文翔，2018）；除了因為範疇相對明確，如此訓練出的模型，在生成語句或段落時，更加接近人所撰寫的文字。

最後透過我們的模型進行預測分類，將這些內容進行語意分類建構正向、負向及中立的類別。在 6,917 筆完成人工標記的語料中，先排除三位標記者互不同意的筆數，將語料隨機切分為 70%和 30%，前者作為訓練資料集(training dataset)，後者做為測試資料集(test dataset)。換句話說，我們的模型以 BERT-Chinese 預訓練模型為基礎，運用我們人工標註的 70%進行該預訓練模型的微調(fine tune)機制，接著，再將剩餘的 30%交由模型預測情感分類，再對比人工標註與模型預測的結果，以評估我們的模型效果。

4. 資料分析結果 (Results)

4.1 標記結果比對 (The Comparison between the Manual Coding and Model prediction)

在測試資料集中，我們將人工標註的結果當作基準真值(ground truth)，與模型預測的類別進行比較（詳見圖 1）。圖 1 顯示各分類之下的貼文數量，其中，藍色（深色）表人工標註的結果，橙色（淺色）表模型預測的結果。對於負向(N)貼文可以看出人工標註的數量大於模型預測的結果，然而，正向(P)和中立(Neu)的貼文數量反而是模型預測的數量大於人工標註的數量；整體而言，無論是人工標註或模型預測的結果，皆是中立(Neu)貼文數量最多。

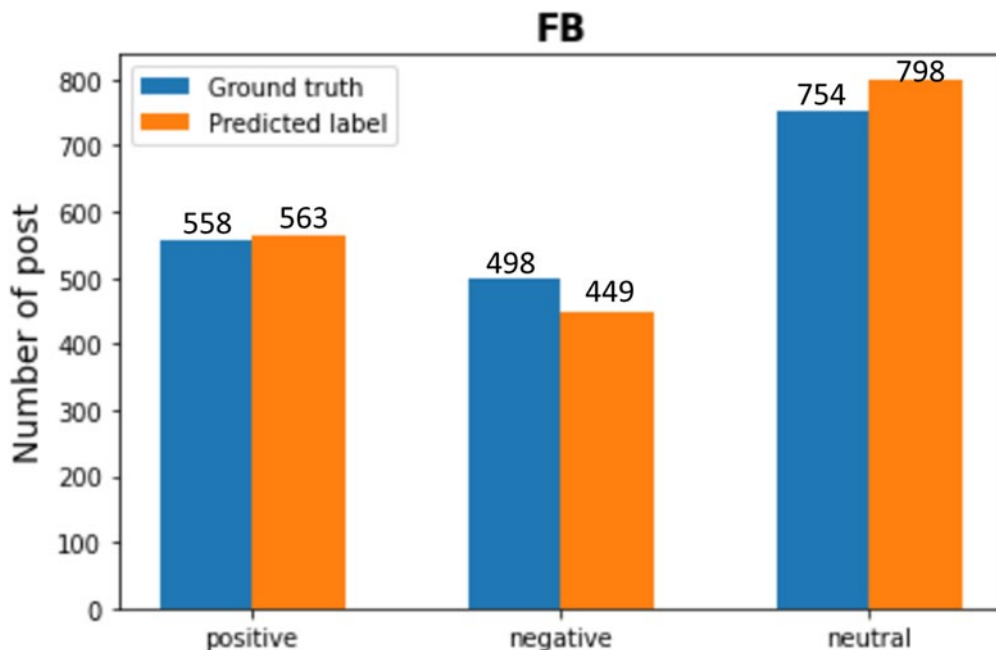


圖 1. 人工標註與模型預測於各類別貼文的數量
 [Figure 1. The number of manually coded posts and model predicted posts

透過圖 1，我們發現人工標註為正向與模型預測為正向的貼文差異最少，人工標記為正向共 558 筆貼文，模型預測為正向共 563 筆貼文，差異數量對應人工標註的比例僅 0.9%³，顯示正向貼文預測效果最好；而人工標註為負向與模型預測為負向的貼文差異最多，差異數量對應人工標註的比例達 9.8%。為進一步探討這些差異的問題，我們利用混淆矩陣(confusion matrix)來評估語意分析的模型（詳見圖 2）。圖 2 縱軸為人工標記，橫軸表模型預測，軸線上的 P 表示正向、N 表示負向，而 Neu 表示中立。透過混淆矩陣可以看出人工標註和模型預測於各類別貼文的正確率及錯誤率。

就正確率而言，圖 2 顯示，在人工標註為正向貼文中，模型預測為正向的比例為 0.64⁴；人工標記為負向的貼文中，模型預測也為負向者為 0.63；人工標記為中立的貼文中，模型預測也為中立者達 0.81，為這三種情感分類正確率最高的數值。

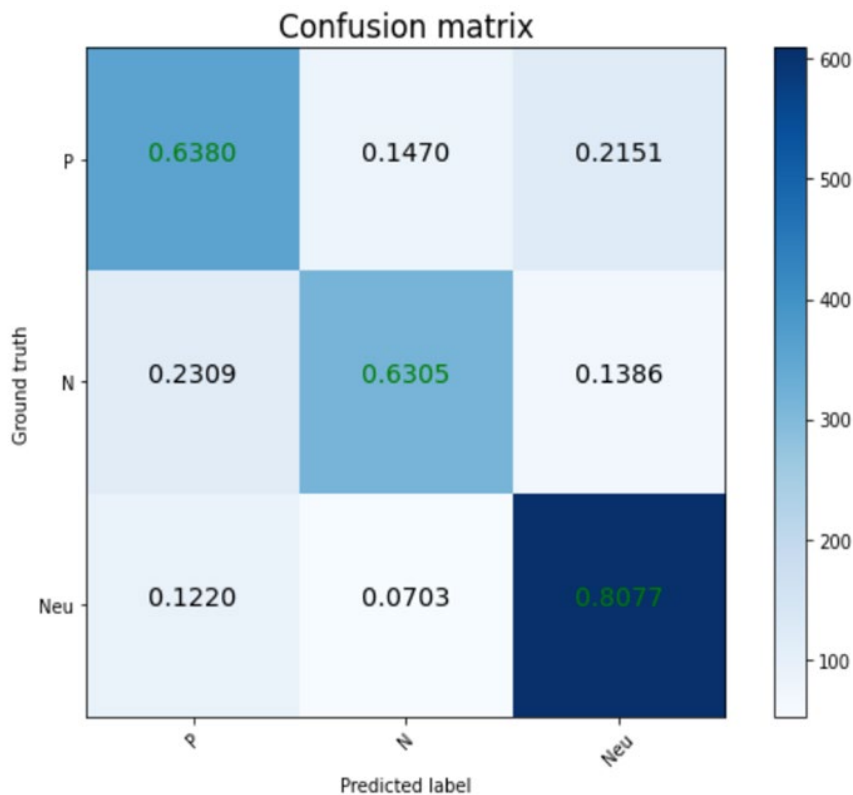


圖 2. 混淆矩陣
[Figure 2. Confusion matrix]

³ 該比例計算方式為： $(563-558)/558=0.896\% \approx 0.9\%$ ，人工標註為負向與模型預測為負向的貼文差異比例計算方式亦同，即 $(498-449)/498=9.83\% \approx 9.8\%$

⁴ 圖 2 的數字我們四捨五入取至小數點後第二位，因此，人工標記為正向而模型預測為正向的 0.6380 我們四捨五入為 0.64，其他數字亦同。

就錯誤率而言，人工標記為正向的貼文中，模型預測為負向或中立的比例分別為 0.15 和 0.22；而人工標記為負向的貼文中，模型預測為正向或中立的比例分別為 0.23 和 0.14；人工標記為中立的貼文中，模型預測為正向或負向的比例分別為 0.12 和 0.07。錯誤率中，人工標記為負向而模型預測為正向者最高，其比例為 0.23；錯誤率最低者為人工標記為中立而模型預測為負向，其比例為 0.07。

綜上所述，人工標記為負向而模型預測也為負向的正確率最低，僅 0.63，其次為人工標記為正向且模型預測也為正向者，0.64，這顯示該模型對於負向和正向情感貼文的判斷能力較差。人工標記為負向的貼文中，有 0.23 模型預測為正向貼文，為所有錯誤率最高者；錯誤率次高者為人工標記為正向而模型預測為中立的 0.22，這兩個現象顯示，人工判斷屬於負向或正向情感的貼文，預訓練模型卻預測為正向或中立情感。這將讓我們進一步探討箇中原因。

4.2 語言特徵 (The Linguistic Features)

圖 2 顯示，人工標記為負向的貼文中，模型預測為正向（以下簡稱「人工負－模型正」）的比例為 0.23，為所有錯誤率最高者，其次為人工標記為正向而模型預測為中立（以下簡稱「人工正－模型中立」）的 0.22。我們因而聚焦於這兩個部分。排除三人標記不一致的筆數，以及第三位標記者再次確認內容所傳達的訊息後認為應調整情感分類的貼文，我們檢視並分類負向貼文和正向貼文的語言特徵，以利為模型後續發展。

我們逐筆檢視，試圖透過人工標註和模型預測之間的差異來歸納特定情感的語言特徵。我們一方面認同 Nissim & Patti (2017) 的看法，認為詞彙層次的分析已不足以判斷貼文的情感，而應納入跨詞訊息（如句法）、非詞訊息（如標點符號和表情符號）及言談訊息（如類比）等；另一方面，我們主張應考量跨欄訊息（如貼文者身分）以及時事知識，方能較完整取得貼文的情感樣貌。

我們雖然設定「校正回歸」為關鍵詞，但基於臉書使用者不同目的和特性，這些語料大部分聚焦於校正回歸相關新聞，但仍有一部分圍繞其他主題，例如股市、旅遊、購物、美食…等，我們一併納入討論。我們先分別討論負向情感和正向情感之語言特徵，然後再比較兩種情感語言特徵的異同。

4.2.1 負向情感語言特徵 (Linguistic Features of Negative Polarity)

我們將負向情感的語言分為（甲）－（庚）類：（甲）貼文使用負向評價語言，（乙）貼文使用標點符號、表情符號或語尾助詞傳達負向情感，（丙）貼文使用類比，（丁）貼文訴諸其他國家，（戊）貼文使用反語或反串文，（己）貼文引述典故、文學作品、電視劇、電影等，（庚）貼文變造口號。此外，我們也發現有些貼文需要跨欄知識或對時事的了解才能正確判斷情感，我們將之歸類為（辛）。我們對這 8 類分別說明、舉例如下，並以底線標示屬於該類型的語言特徵。

(甲) 貼文使用負向評價語言：這一類貼文使用負向評價語言(施孟賢等, 2021)表達負向情感, 參考施孟賢等(2021), 我們認為正向語言可以是單詞、片語或子句。由於負向語言有不同的形態, 我們又分為下列(甲1) – (甲6)等次類:

(甲1) 使用指責、批評、否定或不信任的語言: 對象可能是政府、政治人物或某些政治人物的支持者, 如【例1】明確批評指揮中心創造新名詞的作法可能使民眾更加驚恐:

【例1】#點新聞 都已經疫情這麼嚴重了, 指揮中心還在那邊 #創造新名詞, 這樣會讓民眾更恐慌吧!

(甲2) 引述、討論名人/專家: 這類貼文多半引述名人/專家批判性的言論, 或報導名人/專家的作為導致負向結果, 或以主題標籤「#」標註名人/專家, 如【例2】引述媒體人李艷秋批評指揮中心作為的言論, 且使用負面言說動詞「開嗆」; 【例3】則直接使用「被罵翻」來說明陳時中提出「校正回歸」一詞所遭遇到的反彈:

【例2】李艷秋開嗆: 辛苦陳時中了, 每天記者會還要提供娛樂表演, 反正大家防疫靠自己, 就當宅在家的消遣吧! [...]⁵

【例3】[...]幕後 | 陳時中「校正回歸」被罵翻 [...]

(甲3) 貼文質疑政府、政治人物或某些政治人物的支持者: 包含一般問句(如【例4】)和修辭性問句(【例5】):

【例4】[...]重點是, 除了 321 例外又有 400 例, 而這 400 例是因為檢驗的時間差造成的, 但為何不直接每天滾動式調整而是一次公佈調整??

【例5】數字可以校正回歸, 但這世界還能校正、回歸到疫情前的生活樣態嗎?

(甲4) 貼文使用貼標籤法(或譯為「咒罵法」): 如同施孟賢等(2021)主張, 標籤可以是人物的標籤, 如【例6】之「巨嬰」, 將小聖蚊和藍白粉(即國民黨與民眾黨的支持者)直接稱呼為「巨嬰」, 表示把這些人「將原本自己該負的責任, 移轉給國家承擔」(呂秋遠, 2018); 標籤也可以是政策或事件的標籤, 如【例7】, 指涉鼓吹政府施打疫苗的倡議為「疫苗業務大會」:

⁵ [...]表示由本文作者刪節的貼文部分; 若是貼文者自行使用的刪節號, 則以...呈現。

【例 6】我真的不希望台灣巨嬰太多 小聖蚊跟藍白粉這些咖小 就跟確診人數一樣、拜託不要再增加了。

【例 7】今天應該不會有人再去討論校正回歸 我想今天最大的擂台賽應該會是「滯台中國人的疫苗業務大會」

(甲 5) 貼文直接使用表達負向情感的文字：包含痛苦、失望、絕望、恐懼、驚嚇、擔憂...等，如【例 8】：

【例 8】恐怖！連 8 日破百！本土確診+321

(甲 6) 貼文表達不理解或猜測：貼文使用表達無法理解或猜測的文字，如【例 9】中的「滿臉問號」：

【例 9】今天學到一個新的詞「校正回歸」，然後大家看著指揮官滿臉問號。

(乙) 貼文使用標點符號、表情符號或語尾助詞傳達負向情感：這一類的貼文，類似 Tannen (2013) 提出的「熱忱標記」，即：使用（有時重複出現的）標點符號、表情符號和語尾助詞等，來強化無奈、驚嚇、崩潰或不信任...等負向情感。標點符號包含刪節號(...)、波浪紋(~)和驚嘆號(!)，【例 10】在表達負向情感的文字「懶得回應」後使用刪節號表達欲言又止的負面情緒，可能是無奈；表情符號包含 QQ、Σ(°Д°;)、😭⁶...等，【例 11】在發現台東縣也出現確診案例後，使用表示流淚的 QQ 加強無奈或絕望等情緒；語尾助詞包含「啊(阿)」、「啦」、「吧」、「耶」等，【例 12】若單純看「台北市的量也太多了」可能難以斷定情緒方向，但在同時使用語尾助詞「吧」和刪節號後，便引發出無奈的負面情感。

【例 10】小編懶得回應了…… 反正大家都知道會有這種「現象」就好了……
P.s. 1.如果真的「來不及」就開始好好講「來不及」，並且有因應作為，而非搞了好幾天，再來一個「校正」，並且甩鍋給地方。[...]

【例 11】台東也淪陷了 QQ

⁶ 手機/Google 表情包的表情符號無法於 Word 檔案中顯示，我們以註腳說明，這個符號是哭臉。

【例 12】#懶編：台北市的量也太多了吧....

(丙) 貼文使用類比：貼文使用隱喻或「情節／劇本(scenario)」(Musolff, 2004, 2006)⁷，表達對於防疫措施或「校正回歸」一詞的負向情感，如「塞車」、「作帳」、「捅一刀」...等，這些隱喻或情節／劇本皆帶有負向含意：【例 13】中的「內外帳」隱喻，來自公司內部帳冊和對外公告帳冊的不一致性，指涉指揮中心對國人隱藏真實確診數字的作為；【例 14】則使用「黑洞」表示大眾無從得知真實的確診數字，指責地方政府防疫採檢通報量的不實；【例 15】使用路人指揮消防員滅火的情節／劇本，來表達對某政黨支持者的不滿。

【例 13】#這是內外帳的差別 台灣統計新冠肺炎確診人數,22 日本土病例 321 例, 境外傳入 2 例, 校正回歸 400 例。 向世衛通報直接大方寫上「+723 人」。 23 日對內宣佈：本土病例 287 例, 境外傳入 3 例, 校正回歸 170 例。 向世衛通報, 是「+460 人」。

【例 14】今天應該關注的重點還是新北的確診數吧, 侯友宜轄下的新北不只是防疫漏洞, 根本就防疫黑洞了

【例 15】要搞鬼或想算帳, 請疫情結束再來, 你們現在就像是火神的眼淚裡面一樣, 兩個消防員因為專業吵就算了, 至少我們知道他們都在努力想辦法解決, 在想辦法找出大方向最好的方法。結果路人還跳出來想指揮消防員滅火, 要朝哪噴水才正確, 也只是想讓人家以為你很厲害而已[...]

(丁) 貼文訴諸其他國家：這類貼文比較台灣和歐美日等國的防疫措施，從而表達對台灣防疫政策的不滿，如【例 16】；或如【例 17】說明他國校正數字並非將確診人數往上加，而是削減確診人數，以糾正台灣對於「校正回歸」的用法：

【例 16】楊艾俐：企業要買快篩有何罪？指揮中心不同意，太不合理了。在美國，甚至個人都可以到藥房去買，別說是企業，在西雅圖的微軟及其他高科技公司早在去年五月，就已經每天快篩要進去的員工，這才降低西雅圖的感染率。美國屠宰場一度傳出有很多確診，後來凡上工的都要在外面快篩，才能進去上班，Amazon

⁷ Musolff (2004, 2006) 主張，情節／劇本為不同場景(scene)或故事線的集結，並且該語言使用者可理解這個情節／劇本中的角色、故事線，以及結果，並可判斷情節／劇本成功與否、正常與否，及／或正當與否等。

也如是，警察、醫護人員都是每天測，現在美國已測試 4 億多劑，平均每人測試 2 次。台灣測 20 餘萬劑，平均 100 人才測試一劑，國王的新衣是該脫下了。

【例 17】和大家分享，其實法國昨天才因應「錯誤重複計算」，「校正」#消除了 35 萬例確診數喔。

(戊) **貼文使用反語或反串文**：根據 Wilson (2017: 202) 的研究，反語(irony)傳統定義為「說一件事，但卻表達反義」。Wilson 認為反語應視為 Sperber & Wilson (1981) 提出的「回響說明(echoic account)」：反語使用者並非說出與字面意義相反的語言，而是針對某一個體或群體，反應出自己的想法，並嘲笑、蔑視這個想法。Wilson (2017: 202) 舉例說明，當一個人說「政客絕不會撒謊」時，他並非斷言政客有時／總是撒謊，而是針對人們對政客懷抱「絕不撒謊」的期待表示輕蔑與嘲諷。Wilson 因而認為，反語的重點並非言談的內容，而是對該內容表達的態度。如**【例 18】**「博大精深」原為褒義詞，義為「廣博深遠」（教育部重編國語辭典修訂本），但出現在「300+400 校正回歸…」之後，凸顯貼文者預期指揮中心對防疫和篩檢應有更佳作為，然而指揮中心卻提出「校正回歸」一詞的失望心態，因而使用反語來表達其態度。再者，這類貼文亦可視為反串文，即貼文者本身並非具有某種身分或立場，但卻假裝自己是該身分立場的人來發言，以達到某種目的。常見的反串包括各種常成為爭吵（戰文）的身分，例如政治立場（藍綠）、支持球隊、性別、學校、職業、南北地域、類組等等。（PTT 鄉民百科）。

【例 18】 300 + 400 校正回歸... 再次說明中文的博大精深

此外，**【例 19】** 中的「真給力」出現在「數字遊戲」之後，同時利用「校」可讀為「ㄒ一ㄠˋ」的特性，造出「笑症回歸」一詞，下接「看好戲」，這首打油詩的貼文者明顯不認同指揮中心提出「校正回歸」的做法，並以反語「真給力」加以嘲諷：

【例 19】 校正回歸四百例，數字遊戲真給力，時中防疫不努力，笑症回歸看好戲。

(己) **貼文引述典故、文學作品、電視劇、電影等**：這類貼文有時直接引述帶有負面意涵的典故或作品，如**【例 20】** 中「國王的新衣」指涉指揮中心在防疫作為上的自我欺瞞，**【例 21】** 的「沐猴而冠」則直指陳時中的牙醫專業無法勝任防疫指揮官一職；有時引述的作品或典故並無特別的正負向意涵，需要透過前後文（如，「笑噴」）來判斷，如**【例 22】**：

【例 20】[...]台灣測 20 餘萬劑，平均 100 人才測試一劑，國王的新衣是該脫下了。
[...]

【例 21】[...]憑恃政治正確的政治人物，沐猴而冠，擔任防疫指揮官，絕無成功抗
疫的可能。[...]

【例 22】台版 TENET 天能 校正回歸成流行語！😂 引爆網友笑噴🔥 有人說蓋牌
就蓋牌嘛，XD！

(庚) 貼文變造口號：疫情初始，政府推出「有政府，請安心」口號⁸，其後演變為「有
政府，好安心」，而【例 23】結合「作帳」隱喻，變造了這個口號，來表達負向情感：

【例 23】果真有政府，會做帳，防疫成果好棒棒。

(辛) 需要跨欄知識或時事認知才能正確判斷的貼文：這類貼文看似表達正向情感，實
際上需要跨欄知識（如，貼文者或粉絲團的政黨立場）或對時事有所認識才能適切判斷
貼文的情感。【例 24】單就貼文內容而言，看起來是善意提醒，然而貼文後方以主題標
籤「#」標註台中市長盧秀燕、她的小編以及「烏龍」一詞。這篇貼文背景為，盧秀燕的
小編在 5 月 22 日「校正回歸」一詞公布後，以盧秀燕的帳號發表「陳時中剛公布的 400
例 校正回歸 什麼鬼？」貼文，引發網友爭議。因此，這份看似善意提醒的留言，實為
挖苦，而這一層的解讀，需要對於貼文者身分和時事的認識方能了解。

【例 24】小編...下次記得換帳號喔🙄 〈#蹲路編〉 #盧秀燕 #直播 #小編 #烏龍

以上為負向情感語言特徵，下個章節我們來探討正向情感語言特徵。

4.2.2 正向情感語言特徵 (Linguistic Features of Positive Polarity)


我們將正向情感的語言分為 (1)–(4)類：(1) 貼文使用正向評價語言，(2) 貼文使用標點
符號、表情符號或語尾助詞傳達正向情感，(3) 貼文使用類比，(4) 貼文訴諸其他國家。
此外，我們也發現有些貼文出現上述語言特徵，但主要目的在促銷商品、服務或節目…
等，即第(5)類以廣告為主要目的的貼文；另有如(辛)的第(6)類，雖然呈現(1)–(4)類語言

⁸ 在蘇貞昌的 YouTube 頻道中，2020 年 1 月 29 日出現標題為「有政府，請安心」的影片，其中
蘇貞昌說「有政府，會做事」以及「有政府，可以放心」，之後轉變為「有政府，好安心」的口
號。

特徵，然而這類貼文實則並非傳達正向情感，需要跨欄知識或對時事的了解才能正確判斷。我們對這 6 類分別說明、舉例如下，並以底線標示屬於該類型的語言特徵。

(I) 貼文使用正向評價語言：這一類貼文使用正向評價語言（施孟賢等，2021）表達正向情感，參考施孟賢等(2021)，我們認為正向語言可以是單詞、片語或子句。由於正向語言有不同的形態，我們又分為下列(1A)-(1F)等次類：

(1A) 使用呼籲團結、鼓勵、肯定積極行動的語言：如【例 25】使用「我們一起確實做好防疫措施，好好照顧彼此」來呼籲人民團結對抗疫情：

【例 25】今日本土 287 例，境外移入 3 例  我們一起確實做好防疫措施，好好照顧彼此 🙏❤️

(1B) 引述、談論名人／專家：引述名人／專家言論，或報導名人／專家的作為導致正向結果，或以主題標籤「#」標註名人／專家，如【例 26】說明政務委員唐鳳的作為，持續改善資料上傳的速度，有效解決以往資料上傳延遲所引發的問題。

【例 26】唐鳳又出手了！先前 500 筆資料要 2 小時、現在只要 10 分鐘，盼能降低地方醫療行政人員的負擔

(1C) 感謝、體諒他人，對他人表達同理心：用於感謝、體諒第一線醫護人員或為疫情提供協助的個人、公司或團體，如【例 27】之「再次感謝所有第一線醫護人員的努力！」，【例 28】則表達對勞工配合防疫作為的感謝：

【例 27】再強調一次，有真實的數據，才能進一步盤點各區情況，才能科學性地研議因應策略、及早部署，讓區域民眾安心。再次感謝所有第一線醫護人員的努力！

【例 28】有些在工地工作的勞工朋友，他們在工作上需要耗費極大的體力，口罩濕了又濕，換了又換，為了防疫，他們還是堅持戴上口罩，亦是最基層的無名英雄，謝謝你們的配合。

(1D) 使用傳達正向情感的口號：這一類的目的亦在呼籲人民團結，並確實做好防疫，但不同於(1A)，這一次類的語言多半為簡潔有力的短語，且一再出現，形成口號，如【例 29】的「台灣加油」：

【例 29】希望是往好的方向走，台灣加油👍👍

(1E) 表示自我肯定的語言：如同 Morin & Flynn (2014)和 Al-Tahmazi (2015)主張，臉書使用者會透過貼文來建構自我或內部團體，並使用正面自我(positive self-presentation, van Dijk, 2006)強化團結，如【例 30】，貼文者以確診數下降的趨勢肯定指揮中心使用「校正回歸」的作為：

【例 30】校正回歸後，我們的確症數趨勢真的逐漸往下了[…]

值得注意的是，(1E) 有時會貶抑他者，即藉由負面他者 (negative other-presentation, van Dijk, 2006)的對比，提升自我或內部團體的正向情感，如【例 31】直接批評韓國瑜的支持者不斷指責指揮中心蓋牌的言論：

【例 31】[…] 中央防疫一起來，成立國家篩檢隊，集合起民間業者力量。#當個堂堂正正的韓粉 #就繼續跳針蓋牌吧。。

(1F) 對於未來的期待與祝福：這個次類表達貼文者的期待與祝福，如【例 32】表達對於返回正常生活的期望。

【例 32】【#TOPick 新聞】希望台灣疫情早日緩和！

(2) 貼文使用標點符號、表情符號或語尾助詞傳達正向情感：這一類貼文使用（有時重複出現的）標點符號（驚嘆號、波浪符號）或表情符號、語尾助詞（「吧」「呀」「啦」…等）來表達驚喜、興奮，或善意提醒等正向情感。【例 33】以「各位都要乖乖待在家裡呀」提醒人民團結對抗疫情；【例 34】則是以「我們一起遵守防疫措施，度過疫情挑戰」加上祈禱和愛心表情符號，以及在「全面戴口罩 #務必配合實聯制，保護好彼此」之後，加入舉臂加油表情符號，呼籲人民團結對抗疫情。

【例 33】各位都要乖乖待在家裡呀（#Y編）

【例 34】今日本土 334 例，境外移入 5 例 📄 我們一起遵守防疫措施，度過疫情挑戰 🙏❤️⁹[⋯] #全面戴口罩 #務必配合實聯制，保護好彼此 🙌🙌¹⁰

(3) 貼文使用類比：正向情感貼文因引述指揮官的發言，使用了「塞車」、「蓋牌」和「作帳」等隱喻，這些隱喻大多帶有負面含意。【例 35】中的「塞車」，說明由於採檢量暴增，許多檢體來不及判讀，一周內便累積了一萬多件舊案延誤時間的情況，讓讀者透過車輛擁擠造成的交通阻塞，來理解採檢量過多而造成回報延誤的情況；「蓋牌」為博弈術語，表示隱藏牌面以迷惑對手；【例 36】的「作帳」意為改變數字以企圖美化帳目。為弱化這些隱喻的負面含意，正向貼文往往在其前後文中，出現修辭性問句（如【例 35】的「這樣的數字算是「#蓋牌」嗎？」）、反義詞（如【例 36】的「掀牌」）或否定詞（如【例 36】的「不是」）：

【例 35】#校正回歸 的問題：(20210522) 會有這樣的問題，因為 #北市 跟 #新北 #篩檢塞車 這兩地被校正的最多、其他地區數字是比較零星。塞車的問題會陸續解決，再行政上面簡化後，會快點出來，這樣子對疫情狀況會有幫助。[⋯]為什麼要採校正數字、主要是疫情評估的判斷、陽性率、趨勢的判斷，會影響疾管署的措施。這樣的數字算是「#蓋牌」嗎？把數字都秀給你了，這樣子叫做蓋牌... 確診數字不是匯市、股市冷冰冰的數字，這些是確診數字，一個數字後面有一個人呢。

【例 36】《單日「721 例」解謎 新增 321 例、400 例「校正回歸」！陳時中：這是掀牌、不是作帳》

(4) 貼文訴諸其他國家：如同負向情感語言中的（丁）類，正向情感也訴諸其他國家，特別是先進國家的作為。不同於（丁）的是，這類貼文的目的是為指揮中心使用「校正回歸」一詞辯護，稱「校正回歸」並非新創名詞，歐美日等先進國家和其他國家早已使用，指揮中心使用這個詞是為了讓確診數字合理，如【例 37】：

【例 37】不是蓋牌！ICU 醫生解釋，「校正回歸」是將處理好的舊通報案件，加回前面的日期，在歐美都有參照。

(5) 貼文以廣告為主要目的：這類貼文可能一部分包含事實報導，也可能利用上述表達

⁹ 這三個表情符號分別是祈禱、方塊和愛心。

¹⁰ 這兩個符號皆為舉臂加油。

正向情感的語言，但通常主要目的為宣傳某團購、餐廳、節目、品牌、政治人物…等，如【例 38】出現第(1A)和(2)類，同時也出現百貨公司名稱進行宣傳：

【例 38】[...]#遠百信義 A13 #防疫日常 🍷 #全民防疫共同守護 ❤️ #安心賣場 🍷 #
防疫人人有責

這一類貼文也有僅宣傳產品，只在文末使用主題標籤#標註「校正回歸」，如【例 39】：

【例 39】[...]🍷不以大集團式的華麗廣告行銷手法，而是回饋實實在在的風土。
🍷裕森老師說酒莊的 Riesling 也深受似乎在這片有紅色石灰質土的園中活得很快活
沒有太多生長的壓力和困難 喝來溫暖而柔和 2 款經典新品以絕佳的品質立足
澳洲酒海中 ❤️ 更是所有酒迷探索年度最佳酒莊的最佳捷徑！ ❤️ #酪洋國際酒
業 #JAMESSUCKLING #JimBarryWines #TheArmagh #Shiraz #CabernetSauvignon #
價錢超優惠 絕不 #校正回歸

(6) 需要跨欄知識或時事認知才能正確判斷的貼文：這類貼文看似表達正向情感，實際上需要跨欄知識（如，貼文者身分）或對時事有所認識才能適切判斷貼文的情感。【例 40】單就貼文內容而言，看起來是使用正向情感詞彙「創造」，然而貼文者為政黨傾向偏國民黨的「李姓中壢選民」，並引用聯合報的民意論壇，故而無法單就貼文內容判斷情感：

【例 40】#校正回歸 #創造新名詞 圖片來源：民意論壇：聯合報。世界日報。udn
tv

在檢視貼文時，我們也發現，有些貼文僅使用單一類別或次類，如【例 41】僅使用（乙）刪節號表示貼文者的無奈；多數貼文則使用一個以上的類別或次類，【例 42】出現了(1B)和(3)：引述了專家（陳時中）的談話內容，而其談話內容則使用了類比，以「蓋牌」比喻隱瞞實際的確診人數，「掀牌」比喻揭露確診人數，「作帳」比喻美化確診人數的數字以圖粉飾太平：

【例 41】想篩檢，只能去排隊 ... #海角天編 【快點 TV】
<https://gotv.ctitv.com.tw/2021/05/1778685.htm>

【例 42】陳時中表示：這是掀牌，怎會是蓋牌？ #校正回歸 #新冠肺炎 《單日「721 例」解謎 新增 321 例、400 例「校正回歸」！陳時中：這是掀牌、不是作帳》[…]

我們根據「人工負－模型正」和「人工正－模型中立」的語言特徵類別的筆數與比例製作表 1 和表 2：

表 1. 負向語言特徵類別

[Table 1. Categories of linguistic features of negative polarity]

負向語言特徵類別		筆數	比例 (%)
貼文使用負 向評價語言 筆數總和 101 比例 51.79%	(甲 1) 使用指責、批評、否定或不信任的語言	34	17.44
	(甲 2) 引述、談論名人／專家	11	5.64
	(甲 3) 貼文質疑政府、政治人物或某些政治人物的支持者	20	10.26
	(甲 4) 貼文使用貼標籤法	10	5.13
	(甲 5) 貼文直接使用表達負向情感的文字	21	10.77
	(甲 6) 貼文表達不理解或猜測	5	2.56
(乙) 貼文使用標點符號、表情符號或語尾助詞傳達負向情感		26	13.33
(丙) 貼文使用類比		12	6.15
(丁) 貼文訴諸其他國家		4	2.05
(戊) 貼文使用反語或反串文		36	18.46
(己) 貼文引述典故、文學作品、電視劇、電影等		5	2.56
(庚) 貼文變造口號		2	1.03
(辛) 需要跨欄知識或時事認知才能正確判斷的貼文		9	4.62
總和		195	100

表 2、正向語言特徵類別

[Table 2. Categories of linguistic features of positive polarity]

語言特徵類別		筆數	比例 (%)
貼文使用正向評價語 筆數總和 65 比例 54.62%	(1A) 使用呼籲團結、鼓勵、肯定積極行動的語言	27	22.69
	(1B) 引述、談論名人／專家	24	20.17
	(1C) 感謝、體諒他人，對他人表達同理心	3	2.52
	(1D) 使用傳達正向情感的口號	3	2.52
	(1E) 表示自我肯定的語言	3	2.52
	(1F) 對於未來的承諾、期待與祝福	5	4.20
(2) 貼文使用標點符號、表情符號或語尾助詞傳達正向情感		25	21.01
(3) 貼文使用類比		5	4.20
(4) 貼文訴諸其他國家		2	1.68
(5) 貼文以廣告為主要目的		20	16.81
(6) 需要跨欄知識或時事認知才能正確判斷的貼文		2	1.68
		119	100

表 1 顯示，第（甲）類「貼文使用負向評價語言」比例最高，共計 101 筆，佔所有筆數 51.79%，而（戊）類「貼文使用反語或反串文」比例次高，共 36 筆，佔 18.46%，第三高比例的語言特徵則為（乙）類「貼文使用標點符號、表情符號或語尾助詞傳達負向情感」，共 26 筆，佔 13.33%，但這三類貼文卻被模型視為正向情感，顯示模型的不足。表 2 顯示，第(1)類「貼文使用正向評價語言」比例最高，共計 65 筆，佔所有筆數 54.62%，次高為第(2)類「貼文使用標點符號、表情符號或語尾助詞傳達正向情感」共 25 筆，佔 21.01%，若排除貼文以廣告為主要目的，第三高則為第(3)類「貼文使用類比」，共 5 筆，佔 4.20%。

4.3 討論 (Discussion)

我們進一步討論負向情感和正向情感的語言特徵類別（含次類）的異同。就相異之處而言，除明顯表達負向或正向情感詞彙的使用（如（甲 1）和(1A)…等），負向情感使用了「貼標籤」、「反語或反串文」、「引述典故、文學作品、電視劇、電影等」以及「變造口號」等，而正向情感使用了「傳達正向情感的口號」。

就相同之處而言，兩種情感皆「引述、談論名人／專家」、「使用標點符號、表情符號或語尾助詞表達負向／正向情感」、「使用類比」、「訴諸其他國家」，以及「需要跨欄知識或時事認知才能正確判斷的貼文」，然而，貼文者的意圖使這些語言特徵產生相當大的差異，我們分別討論如下：

4.3.1 引述、談論名人／專家 (Posts Citing or Discussing Celebrities/Experts)

雖然同樣「引述、談論名人／專家」，使用的詞彙和內容卻迥然不同：「人工負一模型正」的貼文者在引述名人／專家談論校正回歸時，選擇帶有負向情感的言說動詞，如「(開) 嗆」、「質疑」、「爆料」、「解密『校正回歸』內幕」，如上文【例 2】；若言說動詞為中性的「表示」或單純使用冒號，引述內容則傳達負向情感，如【例 43】的「自滿」、「膨風」和「疫情肆虐」，與【例 44】的「有何罪」和「不合理」：

【例 43】「書記長鄭麗文表示，國際媒體一致定調台灣是自滿加膨風，才造成疫情肆虐，證明民進黨若繼續大內宣，無法戰勝病毒，請蔡政府停止政治防疫，進入科學務實的防疫[...]

【例 44】[...]「楊艾俐：企業要買快篩有何罪？指揮中心不同意，太不合理了。」[...]

反觀「人工正一模型中立」的貼文在引述、談論名人時，言說動詞通常使用中性的「說」、「表示」、「指出」、「列出」、「解釋」、「說明」、「講解」、「公布」、「宣布」，甚至是帶有正向情感的「說情」，且引述內容通常在解釋緣由或澄清疑慮，或者帶有正向情感，如【例 45】；也有不使用言說動詞，僅使用引述內容傳達正向情感再加上主題標籤 #，如【例 46】

【例 45】市長柯文哲特別為大眾講解，採檢完到確診一般會有兩天的時間差，但最近檢體數量太多，「時間差」就會超過 2 天，因而出現校正回歸的情況。侯友宜指出，因為時間、資訊有落差，往往中央在確診個案篩檢過程中，有擁擠的狀況，時間有落差，光新北市最高達落差 6 天，為了防止造成防疫破口，中央防疫一起來，成立國家篩檢隊，集合起民間業者力量。

【例 46】校正回歸是要回歸到疫情的「真實面」，目前也積極解決問題中！ #校正回歸 #陳時中

4.3.2 使用標點符號、表情符號或語尾助詞表達負向／正向情感 (Posts Showing Negative or Positive Polarity with Punctuation Marks, Emoticons, or Sentence Final Particles)

負向／正向情感的貼文皆使用標點符號、表情符號或語尾助詞，其中，表情符號包含英文字母結構，例如 QQ 或 XD，或特殊符號和標點符號結合之 Σ(°Д°;)，以及手機表情包

中的符號，如哭臉、舉臂加油等，以明確表達特定情感。語尾助詞因有多樣功能，例如，「啦」同時具有興奮和不耐等兩種情感，仍須視前後文而決定。

標點符號方面，兩種情感皆使用波浪紋和驚嘆號，但我們發現，刪節號(…)為負向情感獨有，我們主張貼文者使用刪節號，表達無奈、失望、難過、擔憂等負向情感。目前的模型並未考量標點符號、表情符號或語尾助詞等，未來在發展模型時建議納入。

4.3.3 使用類比 (Posts Employing Analogy)

兩種情感使用的類比包含了隱喻及情節／劇本(scenario)，且大多帶有負面含意，例如「塞車」、「蓋牌」、「捅一刀」等。負向情感貼文除了上述隱喻，尚有「黑洞」、「作帳」、變造防疫指揮官名字的「時鐘一直都不準」，以及路人指揮消防員救火事宜之情節／劇本；正向情感貼文因引述陳時中的發言，無可避免地使用了「塞車」、「蓋牌」、「作帳」等隱喻，在使用這些帶有負面含意的隱喻時，正向情感貼文會使用修辭性問句、反義詞或否定詞以弱化其負面含意，而負向情感貼文則直接使用這些帶有負向情感的類比。

4.3.4 訴諸其他國家 (Posts ASpealing to other Countries)

負向／正向情感皆訴諸其他國家，然而用法卻不同，負向情感的貼文者認為歐美日等國的防疫政策較能有效防疫，從而表達對台灣防疫政策的不滿，如前文的【例 16】，正向情感的貼文者則藉由說明其他國家也使用「校正回歸」來正當化台灣指揮中心使用「校正回歸」的舉措，如前文的【例 37】。

4.3.5 需要跨欄知識或時事認知才能正確判斷的貼文 (Posts Requiring the Knowledge of News and of the Facebook Account Owner's Political/Social Inclination for Correct Prediction)

兩種情感皆有這類貼文，凸顯有些貼文無法單就內容來正確判斷貼文的情感面向，需要進一步了解貼文者身分、政黨傾向，有時甚至需要具備時事知識，方能正確判斷。在未來發展模型時，宜將貼文者的身分納入考量。

最後，我們也發現「貼文以廣告為主要目的」僅出現在「人工正一模型中立」的貼文中，且佔總所有語言特徵的 16.81%，顯示臉書的廣告文通常訴求正向情感。

5. 結論 (Conclusion)

本研究搜集台灣社群媒體上的使用者（包括社群媒體小編、一般社群媒體使用者）生產資料為分析對象，嘗試以既有之中文自然語言處理模型進行分析，並與人工標記結果進行比對，以了解模型預測與基準真值之間的落差，及其形成的可能原因。由此，本研究可以提出調校繁體中文自然語言處理模型面對社群媒體資料時的可能路徑。

我們的研究結果已回答了第 1 章提出的研究問題。首先，在人工標註和模型預測之差異方面，本研究的結果顯示，我們的模型採用 BERT-Chinese 預訓練模型作為模型參數

的初始值，再運用我們的訓練資料集進行模型訓練及微調機制，進而預測測試資料集貼文內容的情感傾向。對於人工標註為中立的貼文，模型有較好的預測能力，正確率高達 0.81；然而，人工標註為正向及負向的貼文，模型預測之正確率分別為 0.64 和 0.63，而其中在「人工負－模型正」的錯誤率最高，達 0.23，其次為「人工正－模型中立」的貼文，錯誤率達 0.22，可以看出當以社群媒體的貼文為語料時，人工標記為負向的貼文往往被模型判斷為正向貼文，而人工標記為正向的貼文，模型往往預測為中立。

為提供未來模型訓練的方向，我們逐筆分析「人工負－模型正」及「人工正－模型中立」的貼文，並分別歸納出 7 類負向情感和 4 類正向情感的語言特徵¹¹，期能讓模型在處理繁體中文之社群媒體貼文時，提升預測的正確率。

本研究發現，若僅聚焦 text/message 欄位，模型很容易將反語或反串文標註為正向或中立（例如：「校正回歸，請造句」），我們因而主張，社群媒體的發文，除了對文本進行分析，亦需考慮發文者本身的政治與社會位置。如同語言學家分析詞彙時常考慮一詞彙前後的共現詞，對社群媒體的分析，也需考慮文本之外的其他線索。這種跨欄位的分析模式，是否有可能寫入自然語言學習的模型中，以便事先考慮其對分析結果的影響？例如，在訓練模型前，先將發文社團或個人的政治傾向加以分類，並在預訓練中將此設為變項，換言之，將貼文者的政黨屬性納入預訓練的一部分，是否能使模型更正確判斷該貼文的情感向度？

從上述的研究成果可以發現，由於社群媒體素材是由眾人一起生產出來的，因此相較於訓練素材取自字典、教科書、專業新聞報導的模型，本研究分析的語料具有文本較短、用字較為口語化（如：以「北車」指稱「台北車站」）、語句不一定完整（如：以空格隱藏明示意義）、重複音節、使用標點符號（如：本研究發現之刪節號為表達負向情感所獨有）或表情符號表達特殊情感、語詞意義層次變化多（如：隱喻、反語）等特質，這些皆可提供模型未來訓練的方向。換言之，社群媒體的語言有別於模型預訓練使用之書面語或正式語言，我們建議未來宜持續發展適合社群媒體語意的預測模型，並將上述語言特徵納入考量，在模型對社群媒體情感向度的判斷上，將助益良多。

本研究的限制，在於我們使用訓練資料集訓練後，模型在負向和正向的情感判斷上仍難符合人工標記的結果，我們認為這是訓練資料集不足所致，在有限的時間下，我們僅能完成這些語料的標記，特別是（戊）類「貼文使用反語或反串文」，模型預測的結果差強人意，這來自於反語和反串文本身的複雜性：Wilson (2017)認為反語是否成立，取決於發言者是否抱持諷刺態度、情境／事件／表現是否違反人們期待的規範，以及發言者是否使用諷刺語氣，而反串文則是貼文者偽裝自己是某種身分立場的人來發言，以達到特定目的。這些條件，若抽離社會、政治脈絡，便難以正確判斷。若期待模型在預測這類貼文有良好表現，則需以大量資料訓練模型，以 NLP 被廣泛使用的 BERT 模型為例，它是由 Wikipedia (25 億字)及 Books Corpus (8 億字)等未經標籤的大量文本資料作為

¹¹ 相關臉書貼文及其語言特徵的類別標記公開於下列網址：
https://www.cs.nccu.edu.tw/~sichiu/Category_0509.xlsx

預訓練模型的資料集；而 BERT-Chinese 的模型是採用約 50 萬字詞進行預訓練模型；換言之，需要以如此大量的資料進行訓練方得以建構預訓練模型。若要處理反語或反串文，亦需要大量的反語和反串文進行訓練，方得以建構能分辨這類語言使用之模型。

致謝 (Acknowledgements)

本論文承蒙科技部計畫 MOST 109-2410-H-004-MY2 補助，諸位匿名審查委員的指導，以及政大傳播學院碩士學位學程楊雅安、王孝成、陳姿樺三位同學協助編碼，特此致謝。

參考文獻 (References)

- Al-Tahmazi, T.H. (2015). The pursuit of power in Iraqi political discourse: Unpacking the construction of sociopolitical communities on Facebook. *Journal of Multicultural Discourse, 10*(2), 163-179. <https://doi.org/10.1080/17447143.2015.1042383>
- Chibuwe, A., & Ureke, O. (2016). 'Political gladiators' on Facebook in Zimbabwe: A discursive analysis of intra-Zimbabwe African National Union - PF cyber wars; Baba Jukwa versus Amai Jukwa. *Media, Culture & Society, 38*(8), 1247-1260. <https://doi.org/10.1177/0163443716671492>
- Cui, Y., Che, W., Liu, T., Qing, B., & Yang Z. (2019). Pre-training with whole word masking for Chinese BERT. In ArXiv preprint arXiv:1906.08101. <https://arxiv.org/abs/1906.08101>.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In ArXiv preprint arXiv:1810.04805. <https://arxiv.org/abs/1810.04805>
- Facebook (2021). *Facebook reports fourth quarter and full year 2020 results*. Retrieved from <https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-Fourth-Quarter-and-Full-Year-2020-Results/default.aspx>
- Jain, P. K., Quamer, W., Saravanan, V., & Pamula, R. (2022). Employing BERT-DCNN with sentic knowledge base for social media sentiment analysis. *Journal of Ambient Intelligence and Humanized Computing, 2022*, 1-13. <https://doi.org/10.1007/s12652-022-03698-z>
- Keen, A. (2007). *The Cult of the Amateur. How today's Internet Is Killing Our Culture*. Crown Business.
- Lu, Y., Pan, J., & Xu, Y. (2021). Public sentiment on Chinese social media during the emergence of COVID-19. *Journal of Quantitative Description: Digital Media, 1*, 1-47. <https://doi.org/10.51685/jqd.2021.013>
- Maíz-Arévalo, C. (2015). Jocular mockery in computer-mediated communication: A contrastive study of a Spanish and English Facebook community. *Journal of Politeness Research, 11*(2), 289-327. <https://doi.org/10.1515/pr-2015-0012>
- Morin, D.T., & Flynn, M.A. (2014). We are the Tea Party!: The use of Facebook as an online political forum for the construction and maintenance of in-group identification during the

- 'GOTV' weekend. *Communication Quarterly*, 62(1), 115-133. <https://doi.org/10.1080/01463373.2013.861500>
- Musolff, A. (2004). *Metaphor and Political Discourse: Analogical Reasoning in Debates about Europe*. Palgrave Macmillan.
- Musolff, A. (2006). Metaphor scenarios in public discourse. *Metaphor and Symbol*, 21(1), 23-38. https://doi.org/10.1207/s15327868ms2101_2
- Nissim, M., & Patti, V. (2017). Semantic aspects in sentiment analysis. In Pozzi, F.A., Fersini, E. Messina, E., & Liu, B. (eds.), *Sentiment Analysis in Social Networks* (pp. 31-48). Elsevier Inc.
- Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., & Harabagiu, S. (2012). Empatweet: Annotating and detecting emotions on twitter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3806-3813.
- Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., ... Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529, 484-489. <https://doi.org/10.1038/nature16961>
- Singh, M., Jakhar, A. K., & Pandey, S. (2021). Sentiment analysis on the impact of coronavirus in social life using the BERT model. *Social Network Analysis and Mining*, 11(1), 1-11. <https://doi.org/10.1007/s13278-021-00737-z>
- Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631-1642.
- Sperber, D., & Wilson, D. (1981). Irony and the use-mention distinction. In P. Cole (ed.) *Radical Pragmatics*. Academic Press. Reprinted in S. Davis (ed.) (1991) *Pragmatics: A Reader* (pp. 550-563). Oxford University Press.
- Tabe, C.A. (2016). Language and humor in Cameroon social media. In Taiwo, R., Odeunmi, A., & Adetunji, A. (eds.) *Analyzing Language and Humor in Online Communication* (pp. 131-163). IGI Global.
- Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 1, 1555-1565. <https://doi.org/10.3115/v1/P14-1146>
- Tannen, D. (2013). The medium is the metamessage: Conversational style in new media interaction. In Tannen, D., & Trester, A.M. (eds.) *Discourse 2.0. Language and New Media* (pp. 99-118). Georgetown University Press.
- van Dijk, T.A. (2006). Ideology and discourse analysis. *Journal of Political Ideologies*, 11(2), 115-140. <https://doi.org/10.1080/13569310600687908>

- Wang, T., Lu, K., Chow, K. P., & Zhu, Q. (2020). COVID-19 sensing: Negative sentiment analysis on social media in China via BERT model. *IEEE Access*, 8, 138162-138169. <https://doi.org/10.1109/ACCESS.2020.3012595>
- Wilson, D. (2017). Irony, hyperbole, jokes and banter. In J. Blochowiak et al. (eds.) *Formal Models in the Study of Language* (pp. 201-219). Springer.
- Ye, Z. (2019). The emergence of expressible agency and irony in today's China: A semantic explanation of the new bèi-construction. *Australian Journal of Linguistics*, 39(1), 57-78. <https://doi.org/10.1080/07268602.2019.1542933>
- PTT 鄉民百科。反串文。
<https://pttpedia.fandom.com/zh/wiki/%E5%8F%8D%E4%B8%B2%E6%96%87>。[PTT Netipedia. Fake post. <https://pttpedia.fandom.com/zh/wiki/%E5%8F%8D%E4%B8%B2%E6%96%87>.]
- 呂秋遠 (2018)。呂秋遠專欄：台灣人患了巨嬰症？蘋果新聞網。
<https://tw.appledaily.com/forum/20180914/JS3WXYWXFOI5XOTWLLN343LGI/>。
[Lu, C.-y. (2018). Chiou-yuan Lu's Column: Taiwanese suffer from Macrosomia (giant baby syndrome)? *Apple Online*. <https://tw.appledaily.com/forum/20180914/JS3WXYWXFOI5XOTWLLN343LGI/>.]
- 施孟賢、段人鳳、鍾曉芳 (2021)。中文新聞文本之宣傳手法標記與分析。中文計算語言學期刊，26(1)，79-104。[Shih, M.-H., Duann, R.-f., & Chung, S.-F. (2021). The analysis and annotation of propaganda techniques in Chinese news texts. *International Journal of Computational Linguistics & Chinese Language Processing*, 26(1), 79-104.]
- 陳韋帆、古倫維 (2018)。中文情感語意分析套件 CSentiPackage 簡介。圖書館學與資訊科學，44(1)，24-41。[https://doi.org/10.6245/JLIS.201804_44\(1\).0002](https://doi.org/10.6245/JLIS.201804_44(1).0002) [Chen, W.-F., & Ku, L.-W. (2018). Introduction to CSentiPackage: Tools for Chinese sentiment analysis. *Journal of Library and Information Science*, 44(1), 24-41. [https://doi.org/10.6245/JLIS.201804_44\(1\).0002](https://doi.org/10.6245/JLIS.201804_44(1).0002)]
- 鍾文翔 (2018)。新聞導言之智能生成。中央大學資訊管理學系碩士論文。[Zhong, W.-X. (2018). Intelligent Generation of News Lead (Mater's Thesis). National Central University.]

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
1F., No. 34, Ln. 3, Sec. 1, Jiuzhuang St., Nankang Dist., Taipei City, 115022, 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
1F., No. 34, Ln. 3, Sec. 1, Jiuzhuang St., Nankang Dist., Taipei City, 115022, Taiwan,
R.O.C.

Tel. : 886-2-2788-1638 Fax : 886-2-2651-9386

E-mail: acclcp@acclcp.org.tw Web Site: <http://www.acclcp.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 :

社團法人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：請至本會網頁填妥入會申請表，填妥後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115022南港區舊莊街一段3巷34號1樓
電話：(02) 2788-1638 傳真：(02) 2651-9386
E-mail：aclclp@aclclp.org.tw 網址：<http://www.aclclp.org.tw>
連絡人：黃琪 小姐、何婉如 小姐

社團法人中華民國計算語言學學會
個人會員入會申請書

會員類別	<input type="checkbox"/> 終身 <input type="checkbox"/> 個人 <input type="checkbox"/> 學生	會員編號	(由本會填寫)	
姓名		性別	出生日期	年 月 日
			身分證號碼	
現職		學歷		
通訊地址	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>			
戶籍地址	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>			
電話		E-Mail		
申請人:			(簽章)	
中華民國 年 月 日				

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115022南港區舊莊街一段3巷34號1樓
 電話：(02) 2788-1638 傳真：(02) 2651-9386
 E-mail：aclclp@aclclp.org.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 Banl: _____

Card No.: _____ - _____ - _____ - _____ Exp. Date: _____ (M/Y)

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

CARD HOLDER SIGNATURE : _____

Address: _____

Tel.: _____ E-mail : _____

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 886-2-2651-9386 or Mail this form to :

Association for Computational Linguistics and Chinese Language Processing
1F., No. 34, Ln. 3, Sec. 1, Jiuzhuang St., Nankang Dist., Taipei City, 115022, Taiwan, R.O.C
E-mail: aclclp@aclclp.org.tw Website: <http://www.aclclp.org.tw>

社團法人中華民國計算語言學學 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\$_____ 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 02-26519386 或郵寄至：

115022台北市南港區舊莊街一段3巷34號1樓 中華民國計算語言學學會 收

E-mail: aclclp@aclclp.org.tw

Publications of the Association for Computational Linguistics and Chinese Language Processing

	<u>Surface</u>	<u>AIR</u> <u>(US&EURP)</u>	<u>AIR</u> <u>(ASIA)</u>	<u>VOLUME</u>	<u>AMOUNT</u>
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 訊息為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 <i>IJCLCLP</i> : 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@aclclp.org.tw

Name (please print): _____ Signature: _____

Fax: _____ E-mail: _____

Address : _____

社團法人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) 2788-1638

聯絡人：黃琪 小姐、何婉如 小姐 E-mail: acclcp@acclcp.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 from 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 Style (<http://www.apastyle.org/>)

No page charges are levied on authors or their institutions.

Online Submission: <https://ijclclp.acclp.org.tw/servlet/SignInHandler>

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

For more information, please email to ijclclp@acclp.org.tw

C Contents

Special Issue Articles:

Corpus Linguistics and Discourse Annotations

Preface: Corpus Linguistics and Discourse Annotations..... i
Siaw-Fong Chung, Rafal Rzepka, and Shih-ping Wang
Guest Editors

Papers

The Uniqueness in Speech: Prosodic Highlights-prompted
Information Content Projection in Continuous Speech..... 1
Helen Kai-yun Chen, and Chiu-yu Tseng

Topic Development and Boundary Cues in Hakka Conversational
Discourse..... 27
Shu-Chuan Tseng, and Hsiao-chien Liu

應用文步分析探究言語行為—以公共政策網路參與平臺提案
文類為例 [A Move Analysis of Communicative Acts in Petition
Text on the Public Policy Participation Network Platform]..... 53
*楊惟婷(Wei-Ting Yang), 謝承諭(Chen-Yu Chester Hsieh),
鍾曉芳(Siaw-Fong Chung)*

An N-gram Approach to Identifying the Chinese Linguistic
Signals for the Problem-Solution Pattern in Annotated Online
Health News..... 75
Chen-Yu Chester Hsieh, and Yu-Yun Chang

Let Me Finish!—Speech Patterns of Interruptions in Chinese: A
Corpus-based Study on Parliamentary Interpellations on Taiwan... 111
Christian Schmid, and Chia-Rung Lu

以社群媒體語言建構深度學習模型：以「校正回歸」為例
[Constructing a Deep Learning Model Using Language in Social
Media: The Case Study of Retrospective Adjustment]..... 153
*段人鳳(Ren-feng Duann), 邱淑怡(Shu-I Chiu),
劉慧雯(Hui-Wen Liu)*