

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between “computation” and “linguistics,” originally established for the term “computational linguistics,” has been replaced by machine-learning-based data science to predict linguistic phenomena.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.