香港中文大學 The Chinese University of Hong Kong

Childhood Bilingualism Research Centre 兒童雙語研究中心

Chinese University of Hong Kong-Peking University Joint Research Centre for Language and Human Complexity 香港中文大學-北京大學 語言與人類複雜系統聯合研究中心

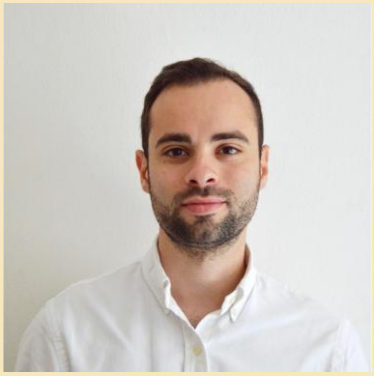
香港中文大學-北京語言大學漢語語言學與應用語言學聯合研究中心 CUHK-BLCU Joint Research Center for Chinese Linguistics and Applied Linguistics

# Building and Analysing YCSEP: The YouTube Corpus of Singapore English Podcasts


## Prof. Alessandro Basile

### Sorbonne Nouvelle University, France


**Date: 4th May, 2026 (Monday)**
**Time: 10:00-11:30 am**
**Venue: Room 220, Fung King Hey Building**



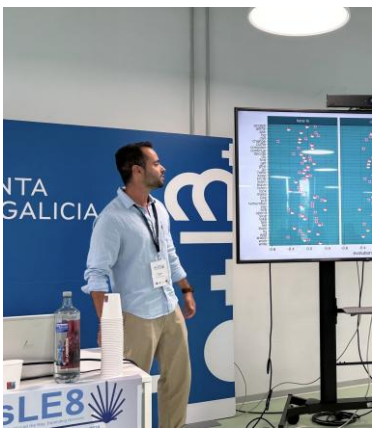
## About This Workshop

This workshop introduces the *YouTube Corpus of Singapore English Podcasts* (YCSEP), a large-scale spoken corpus comprising approximately 620 hours of transcribed, diarized speech drawn from over 1,300 podcast episodes by Singapore-based content creators (Coats et al. 2025). The workshop is structured around two complementary goals: first, a methodological walkthrough of how YCSEP was constructed using a fine-tuned automatic speech recognition (ASR) model; and second, a demonstration of the kinds of linguistic analysis the corpus makes possible.

A central focus of the workshop is the role of ASR in corpus design. Because ASR systems are typically trained on standard, edited speech, they tend to normalize non-standard features – silently 'correcting' them towards mainstream English. Drawing on the development of YCSEP_v2 (Coats et al., *under review*), participants will see how fine-tuning Whisper models on Singapore English data – specifically trained on conversational data from the National Speech Corpus of Singapore (Koh et al., 2019) – substantially improves the recovery of features that baseline systems suppress, including discourse particles and non-standard morphosyntax. It is shown that while baseline models systematically standardize well-documented features of Singapore English such as past tense zero marking or third-person singular *-s* absence (Gupta, 1994; Leimgruber, 2013), the fine-tuned model generally renders them faithfully.

The workshop also highlights the value of this new dataset for the study of contemporary Singapore English, a dynamic variety that has frequently been analysed using datasets collected in the 1990s, such as the *International Corpus of English*. Particular attention will be given to the use of modal constructions of necessity and obligation (Bao, 2010; Basile, 2024), morphosyntactic variables such as copula deletion, aspectual uses of *already* (Bao 2015), and discourse features including sentence-final particles. Participants will gain both insight into building spoken corpora from online sources and a deeper understanding of the linguistic dynamics of a contact variety of English.

## Speaker Bio



**Alessandro Basile** is Associate Professor of English Linguistics at Université Sorbonne Nouvelle. His research interests include grammaticalization, modality, world Englishes, and language contact. He was awarded a PhD in linguistics in 2023 (Paris Cité University) with a doctoral thesis on modal constructions in Singapore English, which forms the basis of his monograph *Modality in Contact: Necessity and Obligation in New Englishes* (2024, De Gruyter Mouton). His research has been published in a range of international peer-reviewed journals, including *Studies in Language, Folia Linguistica Historica, World Englishes, English World-Wide, and English Language and Linguistics*. In 2025, he co-edited the special issue *Cognitive approaches to variation and change in the English modal domain (English Language and Linguistics, 29.3)* and contributed a chapter on constructional change to *The Cambridge Encyclopaedia of Cognitive Linguistics* (CUP). He can be contacted at alessandro.basile@sorbonne-nouvelle.fr .