THE HONG KONG CANTONESE CORPUS: DESIGN AND USES

Kang Kwong Luke*
Nanyang Technological University, Singapore
May L.Y. Wong
The University of Hong Kong

©The Chinese University Press Copyrighted Materials

ABSTRACT

The Hong Kong Cantonese Corpus (HKCC) was built with the specific aim of making available to researchers and language learners a body of naturally occurring talk gleaned from everyday conversations between speakers of Cantonese in Hong Kong.[1] In this paper, we describe the origin, rationale, design principles and uses of HKCC. In particular, we focus on the following aspects of the corpus: (1) data collection procedures; (2) transcription and orthographic conventions; (3) encoding schemes; (4) segmentation and POS tagging; and (5) potential uses of the corpus and future directions.

KEYWORDS

Speech corpus **C**onversation **C**antonese **N**aturally occurring talk **C**orpus design

香港粵語語料庫的設計和用途

陸鏡光 **王麗賢**
南洋理工大學，新加坡 香港大學

提要
建構香港粵語語料庫，旨在爲語言研究及粵語學習提供日常會話中出現的自然語言材料。本文介紹香港粵語語料庫的構思、動機、設計和應用。討論範圍包括：（1）語料收集的原則和過程，（2）轉寫規則，（3）代碼系統，（4）分詞與詞性標注，（5）語料庫的應用及未來發展方向等。

關鍵詞
口語語料庫 **日**常會話 **粵**語 **自**然語言材料 **語**料庫設計

The Chinese University Press Copyrighted Materials