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. 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 Conversation Cantonese Naturally occurring talk Corpus design

1. INTRODUCTION
At a time when language corpora are becoming something of a commonplace in Linguistics, it is perhaps worth reminding ourselves that the use of corpus data has not always been held in high regard. In fact, the idea
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，旨在為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）分詞與詞性標註，（5）語料庫的應用及未來发展方向等。

關鍵詞
□語語料庫 日常會話 粵語 自然語言材料 語料庫設計