A Computerized Lexical Database of Cantonese

Michael C. W. YIP

School of Arts & Social Sciences, The Open University of Hong Kong

myip@ouhk.edu.hk

Introduction

Lexical databases are now available for many languages over the world, for example, CELEX and Euro WordNet. However, there are not any comparable data of this kind for Cantonese. The necessity of this database cannot be underestimated since not much progress on the Chinese (Cantonese) psycholinguistic research can be made without the succor of these kinds of precise lexical information (Li & Yip, 1996, 1998; Yip, 2000). This project aims at making a large-scale collection of digital tape recordings of Cantonese speech and establishing an archive of Cantonese texts based on transcriptions of these recordings. A corpus of Cantonese syllables and words together with other polysyllabic Chinese expressions will be constructed. In the database, much useful lexical information can be generated: for example, spoken word frequency, frequency information of phoneme occurrence and phoneme co-occurrences, speech errors. In this project, a large-scale lexical database of Cantonese will be established in two phases. In the first phase, the target is to have a database of 300,000 Cantonese words. In the second phase, the target will be up to one million Cantonese words. It is hoped that the subsequent psycholinguistics research in the Chinese language can benefit from this computerized lexical database.

Main Objectives

Three main objectives of the project here are:

(1) Generation of relevant lexical information of Cantonese Chinese speech: the database can generate such kinds of useful lexical information for psycholinguistic research as (a) the frequency information of spoken Cantonese word (cf. Yip, 2001); (b) the probabilistic phonotactic information of Cantonese speech (Yip, 2000)

(2) Determination of the processing and production unit of Cantonese speech: from the data of speech error collected in the database, we can closely monitor if the speech error of Cantonese involved a whole syllable replacement or other sub-syllabic components interchanging (cf. Chen, 2000), and then inferred to the functional units of Cantonese speech (Chen & Yip, 2001; Yip, Song, & Chen 1999)

(3) Estimation of the code-switched situation in Hong Kong: from the database, we can estimate the size of code-switching and types of code-switchers in the bilingual situation of Hong Kong (Chan, 1992)

Methodology

This project is designed to construct a computerized lexical database of Cantonese. The database can be used to generate several different kinds of lexical information of Cantonese natural speech. It is based on the large-scale collection of digital tape recordings of natural Cantonese speech. Sources of the natural Cantonese speech include dialogues of Radio call-in programs (Chen, 2000), conversations of TV programs, casual chatting among the students in canteen. Collecting the Cantonese speech from different sources of naturalistic settings guaranteed the ecological validity of the lexical information generated from the database. Because the data gathered to the database is entirely came from the real and natural cases which obviously are psychologically real as well as can reflect the lexical information embedded in our mental lexicon.

Expected Results

The result of this project will be summarized in a computerized lexical database of Cantonese that is significant to research as well as to language teaching and learning. In terms of research, we believe that a more solid rigorous set of lexical information of Cantonese speech can be derived and it can have a wide range of applications to linguistic as well as psycholinguistic researches, especially lexical research centering on spoken language processing. In terms of language pedagogy, it will provide empirical ground for designing the most appropriate language learning methods to students according to the patterns of the prominent processing and production units of native Cantonese speakers. Meanwhile, it will also provide useful information of the pervasive code-switching situation in Hong Kong that clearly confounded the traditional language teaching methods in Hong Kong education sector.

References

Chan, H.-S. (1992) *Code-mixing in Hong Kong Cantonese-English Bilinguals: Constraints and Processes*. MA thesis, Chinese University of Hong Kong.

Chen, H.-C. & Yip, M. (2001). Processing Syllabic and Sub-syllabic information in Cantonese. *Journal of Psychology in Chinese Societies, 2*, 199-210.

Chen J. -Y. (2000) Syllable errors from Naturalistic Slips of the Tongue in Mandarin Chinese, *Psychologia*, 15-26.

Li, P., & Yip, M. (1996) Lexical Ambiguity and context effects in spoken word recognition: Evidence from Chinese. In G. Cottrell. (ed.). *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society.* (pp. 228-232). Mahwah, New Jersey: Lawrence Erlbaum.

Li, P., & Yip, M. (1998) Context effects and the processing of spoken homophones. *Reading and Writing, 10,* 223-243.

Yip, M. (2000) Recognition of Spoken words in continuous speech: Effects of transitional probability. In B. Yuan, T. Huang, & X. Tang. (Eds.), *Proceedings of the ICSLP'2000,* 758-761. Beijing: China Military Friendship Publish.

Yip, M. (2001) A preliminary study of subjective frequency estimates of spoken-words in Cantonese. *Psychological Reports, 88,* 1253-1258.

Yip, M., Song, H., & Chen, H. -C. (December, 1999) Cognitive Processing of Speech: The case of Chinese. Paper presented at the International Language in Education Conference, Chinese University of Hong Kong.