Studies on Cognitive Variation in the Chinese Language Across the Taiwan Straits

Benjamin K. TSOU [1*], Andy C. CHIN [2]
Research Centre on Linguistics & Language Information Sciences,
The Hong Kong Institute of Education
btsou@ied.edu.hk [1*], andychin@ied.edu.hk [2]

This paper focuses on linguistic variations across the Taiwan Straits based on a rigorous comparative analysis within a gigantic and synchronous corpus, LIVAC, which has been cultivated over the last 16 years.

LIVAC (http://www.livac.org) has obtained 1.5 million word types by accumulatively analyzing more than 400 million Chinese characters of newspaper texts in six major Chinese speech communities. It is interesting to note that this 400 million character corpus is based on only about 8,000 character types. Furthermore, there is significant lexical variation among Beijing, Taiwan and Hong Kong. It is found that over the past 16 years, only about 90,000 lexical items are commonly found in all three communities, each of which has made use of more than 300k to 400k lexical items. This approach of finding common lexical items however does not take into account the significant variations in the distribution of the items among the three communities. For example, a word occurring only once in a community but mostly in other communities might arguably *not* be considered a word common to all three communities. A rigorous method is thus called for in studying lexical variations across the Chinese communities.

In addition to tracking linguistic change across Chinese communities, the 16-year history of the LIVAC corpus can also be used as a *Monitoring Corpus* by capitalizing on its *synchronous* nature and *homothematic* content, and taking on an innovative Windows approach to track and conduct unusual and meaningful content analysis of salient cultural items. Two examples involving content rich words relating to *BAR* (吧) and *VEHICLE* (車) and their differential derivative developments in the Cross-Taiwan Straits context will be provided. We will examine relevant lexical variations and innovations as well as account for the lack of reciprocity in *Mutual Intelligibility* **(**互懂度**)** among the speech communities of Hong Kong, Taiwan and Beijing.