On the Non-uniqueness of Error Tagging Solutions of Global Cantonese Learner Corpus

Ken S.K. Cheng
The Hong Kong Polytechnic University, Hong Kong
Ho Ka-wai
The Hong Kong Polytechnic University, Hong Kong

Abstract:

This paper reports the latest development of the Global Cantonese Learner Corpus by Chinese Language Centre at the Hong Kong Polytechnic University, with the goal of improving Cantonese instruction by gathering and evaluating data from students with various linguistic origins. This paper will focus on the issue of error tagging, a specific characteristic that differentiates a learner corpus from corpora for other purposes.

In particular, the standardization of error tagging presents a significant challenge due to the current lack of a unified system for annotating linguistic errors. The paper will first provide a review of the earliest error tagging system for Chinese - the "HSK Dynamic Composition Corpus", in which approximately 50 types of error tags are identified. On the other hand, the "Interlanguage Corpus for Non-Chinese Speaking (NCS) Students in Hong Kong" also includes error tags, with special reference to errors commonly made by non-native Chinese primary students in Hong Kong. These annotation systems are valuable references, but both corpora primarily collect written materials such as compositions, worksheets, and homework. Consequently, their error tagging systems do not include markers for spoken language errors, such as phonetic errors.

Currently, the error tagging system most relevant to our corpus is the "Simon Fraser University Speech Error Database (SFUSED) Cantonese," as it is designed for spoken language and has been adapted for Chinese dialects. This system initially categorizes errors into three major types based on linguistic units: “sound”, “morpheme”, and “lexical”, each with subcategories of "substitution," "addition," "deletion," and "blend" errors, such as "phonetic substitution," "morpheme mis-ordering," and "lexical blending." However, a key distinction between this corpus and the SFUSED corpus is that the latter collects errors from native Cantonese speakers, defining the errors they study as slips of the tongue, where the speaker knows the correct form but unintentionally deviates from linguistic norms (Alderete 2024). This differs from the nature of errors made by second language learners. While second language learners do not make errors intentionally, if a learner exhibits systematic errors, these can be considered "habitual deviations" or even signs of fossilization. Given these differences, the analysis and classification of errors cannot be directly transferred, necessitating the development of an error tagging system suitable for Cantonese second language learners.

Nonetheless, even with a unified error tagging system, the specific analysis of errors within sentences can still be non-unique. Consider the following example in the corpus (Cantonese romanization – Chinese characters – English translation):