

Constituency, sequentiality, and prosody: A corpus-based analysis of phraseological sequences in Mandarin spontaneous speech production

Alvin Cheng-Hsien Chen
National Taiwan Normal University
alvinchen@ntnu.edu.tw

Keywords: Multiword sequences, phraseology, prosody, spontaneous speech, usage-based

This study examines the prosodic realization of phraseological sequences in spontaneous speech production of Taiwan Mandarin. Phraseology is defined as the recurrent multiword sequence (RMS) identified in a large corpus. This study used the Mandarin Conversational Dialogue Corpus (MCDC) as a reference corpus to generate a list of RMS. We identified the shared RMS used in the Sinica Phone-aligned Chinese Conversational Speech Database (SPCCSD) to investigate the articulatory characteristics of these RMS. In particular, we investigate whether the prosodic structures of the RMS correlate with their formulaicity and/or syntactic configurations.

To determine the RMS list, recurrent three-word to five-word sequences in MCDC were extracted. We used a directional measure, Delta P (DP) (Ellis, 2006; Gries, 2013) to assess the “glueyness” of the sub-units for each RMS. DP measured a proficient speaker’s likelihood to select a word given a neighboring context of a variable size as a cue. This contingency-based metric allowed us to assess RMS formulaicity in two directions. Cut-off values for the raw frequency and dispersion of the RMS were used to exclude hapax legomena sequences. All RMSs were then ranked by the DP-based “glueyness”. On the other hand, we defined a series of variables to characterize RMS syntactic and prosodic properties. The syntactic properties included (1) the phrasal cohesiveness of the sequence (i.e., whether the sequence is structurally complete), and (2) the (intended) structural type of the sequence (i.e., NP-based, Clause/VP-based, PP-based). Prosodic properties included: co-occurrence with pausing, alignment with intonation units (IUs), and durational reduction. Linear mixed-effect models were used to analyze the relationships between formulaicity scores and syntactic properties of the RMSs on the one hand, and each prosodic property on the other.

Our results suggest that the higher the formulaicity, the more likely the RMS will be aligned with pauses and IUs, and the average syllable duration of the RMS will be shorter. The former prosodic behavior suggests a holistic processing of the RMS and the latter instantiates the common acoustic correlates of frequency effects (Bybee, 2002). A significant interaction between formulaicity and syntactic configuration is also observed. That is, syntactic integrity will have little/strong effect on the prosodic realization of the RMS if the RMS is high/low in formulaicity. This study provides support from speech production for the claim shared by usage-based linguists that constituency is not a fixed categorical attribute of linguistic units, but an emergent property motivated by language use.

References

- Bybee, J. (2002). Sequentiality as the basis of constituent structure. In T. Givón & B. F. Malle (Eds.), *The evolution of language out of prelanguage* (pp. 109-134). Amsterdam: John Benjamins.
- Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, 27, 1-24. doi:10.1093/applin/ami038
- Gries, S. T. (2013). 50-something years of work on collocations. *International Journal of Corpus Linguistics*, 18, 137-165. doi:10.1075/ijcl.18.1.09gri