A new tool for concordancing the Web as a multimodal corpus

Phoebe Lin (The Hong Kong Polytechnic University, Hong Kong)

Formulaic sequences are an umbrella term for conventionalised sequences of two or more words that form holistic functional, meaning and/or processing units. Common types of formulaic sequences, such as idioms, proverbs, speech formulas, and collocations, often display qualities of formal fixedness, semantic non-compositionality, and high frequency of occurrence to different extents. Over the last decade, however, there is a growing interest in the prosodic fixedness of formulaic sequences, including their tendency to align with pauses (Lin & Adolphs, 2009; Erman, 2007; Wray, 2004), their fast speech rate compared with non-conventionalised word sequences (Lin, 2010) and their restricted tonal patterns (Ashby, 2006; Lin, 2013).

The lack of spoken corpora with prosodic annotation or alignment with audio/video streams, however, has hindered the growth of research on the prosodic patterns of formulaic sequences (Lin & Chen, forthcoming). This paper presents a new computer tool developed recently to tackle the shortage of multimodal data and offer researchers a new tool for exploiting the Web as multimodal corpus. Using the online interface, users may compile and concordance their own large, sustainable and dynamic *YouTube* corpora. While the present paper demonstrates the use of the tool for profiling the prosodic patterns of formulaic sequences, the tool can support any research into the interfaces between lexis, prosody and gestures in naturally occurring spoken discourse.

References

Ashby, M. (2006). Prosody and idioms in English. *Journal of Pragmatics, 38*(10), 1580-1597.

Erman, B. (2007). Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics, 12*(1), 25-53.

Lin, P. (2010). *The prosody of formulaic language.* Unpublished doctoral dissertation, University of Nottingham, Nottingham, UK.

Lin, P. (2013). The prosody of idiomatic expressions in the IBM/Lancaster Spoken English Corpus. *International Journal of Corpus Linguistics, 18*(4), 561-588.

Lin, P., & Adolphs, S. (2009). Sound evidence: Phraseological units in spoken corpora. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 34-48). Basingstoke: Palgrave Macmillan.

Lin, P., & Chen, Y. (accepted and forthcoming). Multimodality I: Speech prosody and gesture. In S. Adolphs, & D. Knight (eds), *Routledge Handbook of English Language and Digital Humanities*. London: Routledge.

Wray, A. (2004). 'Here's one I prepared earlier': Formulaic language learning on television. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 249-268). Amsterdam; Philadelphia: John Benjamins.