# Identifying formulaic sequences using transitional probability: A corpus-driven study

Randy Appel & Pavel Trofimovich
Concordia University, Montreal

Formulaic sequences (FSs) are frequently occurring prefabricated multi-word structures, such as *on the other hand* or *the fact that*. From the perspective of usage-based models of language learning, which claim that learning unfolds as meaningful and recurrent patterns are detected in input (Lieven & Tomasello, 2008), FSs emerge as ideal 'building blocks' of language, due to their saliency and frequency of occurrence. Although FSs may offer potential benefits for language learners (Nattinger & DeCarrico, 2001), they are difficult to identify objectively, and current corpus-driven methods yield structurally incomplete, overlapping, or overly extended structures (e.g., *on the other hand the* or *the other hand the*). These can be misleading and are often of little help to language learners or teachers.

This study addressed the challenge of identifying FSs by using transitional probability – a previously unused measure in this field. Transitional probability is a directional measure of word association which can be used to indicate utterance boundaries, thereby leading to more accurate FS identification. As a test case of this statistic, the British National Corpus was used to extract 100 four-item sequences varying in transitional probability between the first three and the final elements (e.g., *for the sake_of* and *more likely to_be*, where '_' designates the transitional probability values of .99 and .32, respectively). The first three elements in each sequence were subsequently presented to 103 native English speakers in a word association task, requiring participants to complete each sequence (e.g., *for the sake*__, *more likely to*__).

Results revealed that the proportion of sequence completions was more closely associated with transitional probability values than with frequency statistics or mutual information indexes, suggesting that many FSs are three-word sequences with a variable slot. Implications of these findings and areas for future research concerning the application of corpus-driven methods to FS identification are discussed.

Word count: 300

Lieven, E., & Tomasello, M. (2008). Children's first language acquisition from a usage-based perspective. In P. Robinson & N. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp.168-196). New York: Routledge.

Nattinger, J., & DeCarrico, J. (2001). *Lexical phrases and language teaching*. Hong Kong: Oxford University Press.