Multiword Expressions – (how) are they universal?

Andreas Buerki
Humboldt-Universität zu Berlin and University of Basel

The phenomenon of multiword expressions (aka formulaic language or multiword units) has in recent years attracted a great deal of research effort in the fields of corpus linguistics, computational linguistics and Natural Language Processing, psycholinguistics and beyond. Multiword expressions (MWEs) are word sequences that recur again and again in largely the same in form language use. They include such items as conversational formulae ('Thank you very much – not at all'), collocations ('face a challenge', 'utter disgrace'), multi-word units ('dual carriage way', 'contempt of court') as well as other usual sequences ('half an hour', 'no chance of X', 'behind closed doors'). MWEs are thought to be of central importance to language in a number of ways. For example, their knowledge is thought necessary for native-like proficiency in a language since MWEs represent usual turns of phrase, a smaller set of expressions than what might be judged grammatical (Pawley & Syder 1983:191, similarly O'Keeffe et al. 2007:60). MWEs are also thought to ease processing load during language production and thus enable fluency (Nattinger & DeCarrico 1992:32; Pawley & Syder 1983; Wray & Perkins 2000) as well as aiding mutual understanding in communication by activating usual situational and cultural background (Feilke, 2003:213). It is generally assumed that MWEs are found in language universally. The points made above arguably predict MWEs not only to be found in all languages but to be found in roughly comparable measure in all languages – it would be difficult to maintain that some languages have more usual ways of expression than others or that fluency and mutual understanding is more easily achieved in some languages than others since they have a larger number of MWEs. To date, however, no quantitative cross-linguistic studies have been carried out to test whether MWEs are indeed found in similar measure in different languages. In this paper, quantitative corpus data taken from three morphologically very different languages (English, German and Korean) will be presented that suggest the assumption of universality is highly problematic: a concept of MWEs as sequences of word-forms, as is frequently employed (e.g. Biber et al. 1999 and subsequent work on lexical bundles), is shown to yield widely differing counts of MWEs in these languages. While a more abstract conception of MWEs therefore appears necessary, it will be suggested that the construction of a cross-linguistically viable concept of MWEs is more challenging than might at first be supposed due in part to the presence of both fairly fixed and more flexible recurrent sequences and the problematic status of the concept of 'word'. Based on results from a pilot study on the basis of a million words of newspaper text in each of the three languages mentioned, some likely components of a cross-linguistic concept of MWEs will be discussed. They include the partial use of lemmas in place of word forms as well as a partial move to the level of morpheme sequences rather than word sequences. A methodological outline for future work will finally be presented that should yield touchstones for a universal concept of MWEs.

References:

Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (eds), *Language and communication* (pp. 191-226). Harlow: Longman

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching.* Cambridge: Cambridge University Press

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching.* Oxford: Oxford University Press

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language and Communication*, *20*(1), 1-28

Feilke, H. (2003). Textroutine, Textsemantik und sprachliches Wissen. In A. Linke, H. Ortner, & P. Portmann-Tselikas (eds), *Sprache und mehr. Ansichten einer Linguistik der sprachlichen Praxis* (pp. 209-230). Tübingen: Niemeyer

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow: Pearson Education