Is a cross-linguistic typology of multiword expressions useful or even possible?

Lars Borin

Baldwin and Kim (2010: 269) provide a "formal definition" of multiword expressions: "lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity". Paradoxically, as the authors themselves recognize, this definition allows for single-word MWEs, a view which may not be shared by all or even most authors. They mention German compounds here, but the definition would arguably apply equally to, e.g., noun incorporation as found in many languages all over the world (Mithun 1984). It also logically allows for MWEs being made up of other MWEs.

In this definition, I take "lexical item" and "lexeme" to be synonymous, and by implication, to mean something like 'lexical word', one of several construals of the term "word" in linguistics, the two other main ones being 'phonological word' and 'grammatical word' (Dixon and Aikhenvald 2002; see also Haspelmath 2011). Although a quite central concern to NLP, the "orthographic word" is not generally accorded much weight in typologically oriented linguistics, as (1) most languages do not have an established written form, and (2) in those that have an established orthography, there may not be word spacing at all, or the word spacing may reflect a mixture of criteria.

If we are interested in formulating a cross-linguistic typology of MWEs, there are at least three – interrelated – kinds of issues that need to be sorted out.

The most central kind of issue is conceptual or terminological, connected to the difficulty of defining what a word is, but the notion of lexeme also turns out to be slippery. This is true for a single language, and even more of an issue in cross-linguistic comparison (Haspelmath 2010). A crucial factor is that the word is treated primarily as a unit of linguistic form, while the lexeme is seen mainly as a content unit. As mentioned above, words are cross-linguistically characterized as grammatical or phonological – i.e., referring to aspects of their form – while lexemes are commonly defined with reference to their (non-derivable or non-compositional) semantics.

The second kind of issue has to do with the limits of lexicalization. Is it possible to determine what kind of meaning will never be represented as a lexical item, and consequently not an MWE (Goddard 2001; von Fintel and Matthewson 2008)?

The third issue concerns the availability of empirical data for cross-linguistic comparison. Typological linguistics works with large language samples, on the order of at least hundreds of languages, aspiring to be genealogically and geographically representative of the languages of the world. This means by necessity that the data drawn upon in typological studies normally consist of secondary language data, i.e., grammatical descriptions of varying degrees of detail, from brief grammatical sketches (typically) to standard reference grammars (rarely). Tailormade questionnaires focusing on specific features are also common in these investigations. Such secondary sources seldom contain information on MWE phenomena. Thus, one of the most used typological databases, the *World Atlas of Language Structures* (WALS; Dryer and Haspelmath 2013), covers close to 200 linguistic features and almost 2,700 languages (although the database as a whole is quite sparse, as most feature-language combinations have no data), but there are no obvious features relevant to MWEs.

Some specific constructions which conform to Balwin and Kim's definition above – but possibly not to all construals of MWEs – have generated a considerable number of publications in typological linguistics, amassing data from many languages. This concerns constructions such as *compounding* (Lieber and Stekauer 2009), *incorporation* (Mithun 1984), *serial verb constructions* (Aikhenvald and Dixon 2006), *light* or *support verb constructions* (Butt 2010). This body of work is obviously relevant to our question.

Can we make any claims about the preponderance of MWEs in specific languages or even propose a typological classification of languages based on this? Jackendoff (1997:156) is often quoted as stating

that the number of MWEs "is of about the same order of magnitude as the single words of the vocabulary". If, contrary to Baldwin and Kim, we were to define MWEs as made up of (orthographical) words rather than lexemes, a logical consequence of Jackendoff's claim would be that languages such as Swedish, Finnish or German, where compounds are written as one orthographic word, would have much less than at least as many MWEs as SWEs, given that compounds make up a sizeable share of the English MWEs (a third of all multiword items in Jackendoff's *Wheel of Fortune* corpus, and about half, if names, titles and quotations are excluded). Further, there are languages such as Kalam (Pawley 1993), with about 100 lexical verb stems (SWEs), and where serial verb constructions abound for meanings where English would have a single verb. At the other end of the spectrum we find the polysynthetic languages, where entire English clauses correspond to a single verb form, possibly containing only one lexical stem (i.e., one lexeme), as in the Eskimo-Aleut languages (Mithun 2009).

Generally against this background: How should we think about cross-linguistic comparability in the domain of MWEs? Are MWEs even meaningfully comparable across languages? How should we weight orthography, phonology, grammar, and meaning wrt each other in such a comparison? What considerations are specific to NLP as opposed to (typological) linguistics? One purpose of this paper is to relate the practical interest in MWEs in language technology, as evidenced by the large body of work published on this topic over the last decade or so, to relevant linguistic – especially typologically oriented – work.

References

Aikhenvald, Alexandra and R.M.W Dixon (eds) 2006. Serial verb constructions: A cross-linguistic typology. Oxford: Oxford University Press.

Butt, Miriam 2010. The light verb jungle: Still hacking away. Mengistu Amberber, Brett Baker and Mark Harvey (eds), *Complex predicates: Cross-linguistic perspectives on event structure*, 48–78. Cambridge: Cambridge University Press.

Dixon, R.M.W and Alexandra Aikhenvald 2002. Word: a typological framework. R.M.W Dixon and Alexandra Aikhenvald (eds), *Word: A cross-linguistic typology*, 1–41. Cambridge: Cambridge University Press.

Dryer, Matthew S. and Martin Haspelmath (eds) 2013. *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://wals.info>, accessed on 2014-12-18.)

von Fintel, Kai and Lisa Matthewson 2008. Universals in semantics. *The Linguistic Review* 25: 139–201.

Goddard, Cliff 2001. Lexico-semantic universals: A critical overview. *Linguistic Typology* 5: 1–65.

Haspelmath, Martin 2010. Comparatve concepts and descriptive categories in crosslinguistic studies. *Language* 86(3): 663–687.

Haspelmath, Martin 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(1): 31–80.

Jackendoff, Ray 1997. *The architecture of the language faculty*. Cambridge, Mass.: MIT Press.

Lieber, Rochelle and Pavol Stekauer (eds) 2009. *The Oxford handbook of compounding*. Oxford: Oxford University Press.

Mithun, Marianne 1984. The evolution of noun incorporation. *Language* 60(4): 847–894.

Mithun, Marianne 2009. Polysynthesis in the Arctic. Marc-Antoine Mahieu and Nicole Tersis (eds), *Variations on polysynthesis: The Eskimo-Aleut languages*, 3–18. Amsterdam: John Benjamins.

Pawley, Andrew 1993. A language which defies description by ordinary means. William A. Foley (ed.), *The role of theory in language description*, 87–129. Berlin: Mouton de Gruyter.