False truths and true falsehoods: Typological linguistics and multi-word expressions

Lars Borin
Nationella språkbanken • Swe-Clarin
Språkbanken Text • University of Gothenburg

Enet-COLLECT WG1 WS, Gothenburg • 6 Dec 2018
Multiword expressions (MWEs) have attracted much attention in NLP over the last decade or so, at least since the publication of Sag et al. (2002).

In general linguistics, the interest in phraseology – which includes the linguistic study of MWEs – goes back much further (see, e.g., Burger et al. 2007).
However, the broad comparative approach characteristic of research in linguistic typology seems not to have played any role in any of this work so far.

Comparative studies of MWEs in NLP (or phraseology in linguistics) have generally been contrastive rather than typological in scope: they deal with a few languages, rather than with a systematic typological sample. (Bakker 2011).

In the case of MWEs, this arguably should be a “variety sample”. Taking the *Ethnologue* (Simons and Fennig 2018) or *Glottolog* (Hammarström et al. 2016) as the basis for genetic classification of the world’s languages, a minimal variety sample should contain ~120/~400 languages, including all the language isolates recognized by the Ethnologue or Glottolog, such as Basque, Kusunda, etc.
The pros and cons of convenience samples

(Photo from <http://aultparksunrise.com/>)

(Placeholder for additional information or context.)
Linguistic typology is broadly concerned with uncovering and formulating generalizations about the limits, distribution and interdependence of various linguistic phenomena in the languages of the world.

It is at heart a data-driven endeavor, relying on data on many and diverse languages in order to cover the full breadth of linguistic diversity.

Adopting a typologically informed view on MWEs raises a number of theoretical and methodological questions.
Some questions about typological properties of MWEs

▶ How are the “words” of MWEs delimited cross-linguistically?
▶ How prevalent are MWEs (in the vocabulary / in texts / across languages)?
▶ What kinds of MWEs are there and how are they distributed (over the vocabulary of a single language / across languages)?
▶ Where do we find comparable cross-linguistic MWE data?
How are the "words" of MWEs delimited?

- Sag et al. (2002: 2) talk explicitly about "idiosyncratic interpretations that cross word boundaries (or spaces)" (emphasis added), but...
- most languages have no orthography, hence no spaces
- Linguists recognize at least three kinds of "words" – grammatical, phonological, and lexical (but rarely orthographic!) (Aikhenvald and Dixon 2002)
- It is doubtful whether we can provide a definition of "word" that will work for all languages. Haspelmath (2011) goes as far as to say that this is not possible at all, at least not for the grammatical word, which would arguably be the strongest candidate for the "W" in "MWE"
Baldwin and Kim (2010: 269, emphasis added) propose a “formal definition” of MWEs, viz.: “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 ‘MWEs’ comprising a single orthographic or phonological word, a view which may not be shared by all or even most authors, but at least Gantar et al. (2018) express a similar view.
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”.

This statement is based on unsystematic data collection from transcripts of the American television game show *Wheel of Fortune* (i.e., spoken/scripted language). However, it is also supported by the corresponding WordNet statistics, where MWEs make up ~40% of the entries (Sag et al. 2002: 2).

Note that if orthography is taken to determine wordhood, languages like Swedish (or Finnish or German) will automatically have about half as many MWEs, because of the way compound nouns are conventionally written.
Then, there are languages such as Kalam, with about 100 lexical verb stems (SWEs), and where it has been claimed that “(m)ore than 90 percent of conventional expressions for actions and processes are phrases or multi-clause expressions” (Pawley 1993: 87).

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; Dorais 2018).

Less extremely, where one language prefers e.g. compounding, another language may use derivational morphology, i.e., the formal addition of non-lexical (i.e., non-word) material: Swedish matsal ‘dining hall’ ~ Finnish ruokala ruoka-la (food-LOCATION)
Are there languages without MWEs? The general view in the literature seems to be that MWEs are universally present in languages.

What is the minimum and maximum share of MWEs in the lexicon of any language?

How diachronically stable are MWEs and MWE types?

How do MWEs behave wrt language contact? (E.g., Aikhenvald (2006: 52) states that “(v)erb serialization as a grammatical mechanism tends to diffuse”)

How can we know?
Jackendoff’s statement cited above concerns the *lexicon* (of English). We would also like to have some information about the *text frequency* of MWEs.

Nivre and Nilsson (2004: 41f) report about 2 MWEs per 100 words of running Swedish text, so:

Why so many MWEs in the lexicon and so few in text?

Cf. Zipf (1935: 25): “The magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences.”

But actually, we simply don’t know the true text distribution of MWEs!
Polinsky (2012) presents research on the correlation between the (lexical) noun-to-verb ratio (N/V) of a language and its word-order typology:

- Basically, head-initial languages display low N/V. Contrary to this, the highest N/V is found among verb-final languages (Polinsky 2012: 353).

- Polinsky (2012: 348) specifically ties her results to the tendency in a language to form light verb constructions – which arguably is more likely if the language is verb-final – rather than, e.g., resort to verb-forming derivational devices in order to make up expressions for name-worthy actions and processes.
What kinds of MWEs are there?

Looking at the linguistic (including lexicographical) and NLP literature, we find at least the following kinds of multi-lexeme expressions being discussed:

- compounds
- collocations (conventionalized syntactic combinations)
- phrasal/particle verbs
- support/light verb constructions (“complex predicates”)
- adpositional phrases
  (English *on edge* ‘nervous; eager’, but probably not Finnish *liemessä lieme-ssä* [broth-INESS] ‘in trouble’)
- polysynthesis and incorporation
  (Mohawk *wa-hi-’sereht.an’yhsko* (PST-3SG>1SG-car-steal) ‘He stole my car’)
- serial verb constructions (“complex predicates”)
  (Bislama *Kali i katem splitem wud.* (Kali 3SG cut split wood) ‘Kali cut the log in two.’)
Linguistic typology works with large language samples (typically secondary or questionnaire data), preferably on the order of at least hundreds of languages, aspiring to be genealogically and geographically representative of the languages of the world.

Such sources seldom contain information on MWE phenomena (see, e.g., Schultze-Berndt 2006: 371ff).

The widely used *World Atlas of Language Structures* (WALS; Dryer and Haspelmath 2013), covers close to 200 linguistic features and almost 2,700 languages, but there are no obvious features relevant to MWEs.

MWE information is absent or very hard to find even in large monolingual reference dictionaries.
Taking stock: Towards a typology of MWEs?

- 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 with respect to each other in such a comparison?
- What considerations are specific to NLP as opposed to (typological) linguistics?
Orthographic words obviously cannot be used in a language-independent characterization of MWEs. We should rather be striving for something similar to Haspelmath’s (2015) definition of serial verbs in terms of “comparative concepts” (Haspelmath 2010). The lexical items making up MWEs could then tentatively be equated with the comparative concept “lexeme” in the sense of “free construct” as defined by Haspelmath (2011: 70).

In other words, the typological enterprise should be to investigate multi-lexeme entities (MLEs) with (some) non-computable properties.

Whether these MLEs are also MWEs will depend on the notion of “word” adopted, which in turn most likely will need to be a language-specific one if it is to be useful in, e.g., NLP or lexicography.
NLP methods for identifying MWE candidates in corpora (see, e.g., Pecina 2010) or even methods for unsupervised word segmentation (e.g., Hewlett and Cohen 2011) could potentially be of great help. In particular, we would expect such approaches to provide tools allowing us to treat conventionalization and lexicalization as gradient rather than categorical phenomena.

Developing a typological methodology relying on primary rather than secondary language data is a strong desideratum in any case. A way of identifying (potential) MWEs in small corpora could in fact be a ‘killer app’ for a new direction in lexical typology (as well as for conventional lexicography) and constitute a large methodological step forward in linguistic typology.
Thank you!


