NATIONELLA SPRÅKBANKEN

GÖTEBORGS UNIVERSITET
INST FÖR SVENSKA SPRÅKET

SWE-CLARIN

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

NATIONELLA SPRÅKBANKEN

Background (1)

SWE-CLARIN

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).

NATIONELLA SPRÅKBANKEN

Background (2)

SWE-CLARIN

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.

NATIONELLA SPRÅKBANKEN

The pros and cons of convenience samples

SWE-CLARIN



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

NATIONELLA SPRÅKBANKEN

The typology of MWEs?

SWE-CLARIN

- 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.



NATIONELLA SPRÅKBANKEN

Some questions about typological properties of MWEs

SWE-CLARIN

- 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?

NATIONELLA SPRÅKBANKEN

How are the "words" of MWEs delimited?

SWE-CLARIN

- 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"

NATIONELLA SPRÅKBANKEN

Multi-lexeme expressions?

SWE-CLARIN

- 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.

NATIONELLA SPRÅKBANKEN

The prevalence of MWEs (1)

SWE-CLARIN

- 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.

NATIONELLA SPRÅKBANKEN

The prevalence of MWEs (2)

SWE-CLARIN

- 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)

NATIONELLA SPRÅKBANKEN

The prevalence of MWEs (3)

SWE-CLARIN

- 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?

NATIONELLA SPRÅKBANKEN

The prevalence of MWEs (4)

SWE-CLARIN

- 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!

NATIONELLA SPRÅKBANKEN

The distribution of MWEs

SWE-CLARIN

- 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.

NATIONELLA SPRÅKBANKEN

What kinds of MWEs are there?

SWE-CLARIN

- 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.anvhsko* (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.')

NATIONELLA SPRÅKBANKEN

Where do we find cross-linguistic MWE data?

SWE-CLARIN

- 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.

NATIONELLA SPRÅKBANKEN

Taking stock: Towards a typology of MWEs?

SWE-CLARIN

- 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?

NATIONELLA SPRÅKBANKEN

MWE = MLE

SWE-CLARIN

- 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.

NATIONELLA SPRÅKBANKEN

NLP to the rescue?

SWE-CLARIN

- 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.

NATIONELLA
SPRÅKBANKEN

Thank you!

SWE-CLARIN



NATIONELLA SPRÅKBANKEN

References

SWE-CLARIN

Aikhenvald, Alexandra 2006. Serial verb constructions in typological perspective. Alexandra Aikhenvald and R.M.W. Dixon (eds), *Serial verb constructions: A cross-linguistic typology*, 1–68. Oxford: Oxford University Press.

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

Bakker, Dik 2011. Language sampling. Jae Jung Song (ed.), *The Oxford handbook of linguistic typology*, 100–127. Oxford: Oxford University Press.

Baldwin, Timothy and Su Nam Kim 2010. Multiword expressions. Nitin Indurkhya and Fred J. Damerau (eds), *Handbook of natural language processing*, 2, 267–292. Boca Raton: Chapman and Hall/CRC.

Burger, Harald, Dmitrij Dobrovol'skij, Peter Kühn and Neal R. Norrick (eds) 2007. *Phraseologie: Ein internationales Handbuch der zeitgenössischen Forschung / Phraseology: An international handbook of contemporary research (2 volumes)*. Berlin: Walter de Gruyter.

Dorais, Louis-Jacques 2018. The lexicon in polysynthetic languages. Michael Fortescue, Marianne Mithun and Nicholas Evans (eds), *The Oxford handbook of polysynthesis*, 135–157. Oxford: Oxford 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 2015-07-24.

NATIONELLA SPRÅKBANKEN

References

II

SWE-CLARIN

Gantar, Polona, Lut Colman, Carla Parra Escartin and Héctor Martinez Alonso 2018. Multiword expressions: Between lexicography and nlp. *International Journal of Lexicography*. doi: 10.1093/ijl/ecy012.

Hammarström, Harald, Robert Forkel, Martin Haspelmath and Sebastian Bank 2016. *Glottolog 2.7*. Jena: Max Planck Institute for the Science of Human History. (Available online at http://glottolog.org, Accessed on 2016-07-10.).

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.

Haspelmath, Martin 2015. The serial verb construction: Comparative concept and cross-linguistic generalizations. *Diversity linguistics: Retrospect and prospects*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at http://www.eva.mpg.de/fileadmin/content_files/linguistics/conferences/2015-diversity-linguistics/Haspelmath_paper.pdf, accessed on 2015-07-24.

Hewlett, Daniel and Paul Cohen 2011. Fully unsupervised word segmentation with BVE and MDL. *Proceedings of ACL-HLT 2011*, 540–545. Portland: ACL.

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

NATIONELLA SPRÅKBANKEN

References

SWE-CLARIN

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.

Nivre, Joakim and Jens Nilsson 2004. Multiword units in syntactic parsing. *MEMURA 2004 – Methodologies and evaluation of multiword units in real-world applications (LREC workshop)*, 39–46. Lisbon: ELRA.

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.

Pecina, Pavel 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation* 44 (1–2): 137–158.

Polinsky, Maria 2012. Headedness, again. Thomas Graf, Denis Paperno, Anna Szabolcsi and Jos Tellings (eds), *Theories of everything. In honor of Ed Keenan*, Volume 17 of *UCLA Working Papers in Linguistics*, 348–359. Los Angeles: UCLA Department of Linguistics.

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger 2002. Multiword expressions: A pain in the neck for NLP. Alexander Gelbukh (ed.), *Computational linguistics and intelligent text processing: Third international conference: Cicling-2002*, 1–15. Berlin: Springer.

NATIONELLA SPRÅKBANKEN

References

IV SWE-CLARIN

Schultze-Berndt, Eva 2006. Taking a closer look at function verbs: Lexicon, grammar, or both? Felix K. Ameka, Alan Dench and Nicholas Evans (eds), *Catching language: The standing challenge of grammar writing*, 359–391. Berlin: Mouton de Gruyter.

Simons, Gary F. and Charles D. Fennig (eds) 2018. *Ethnologue: Languages of the world*. 21st. Dallas: SIL International. Online version: http://www.ethnologue.com.

Zipf, George Kingsley 1935. *The psycho-biology of language*. Boston: Houghton Mifflin.