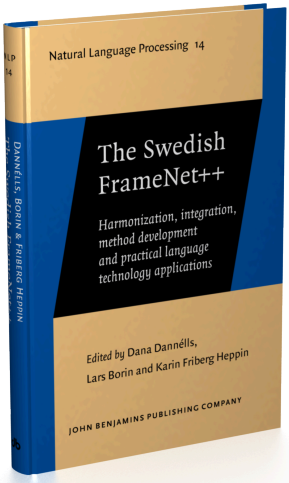
Natural Language Processing 14

CHAPTER 9

Multiword expressions – a tough typological nut for Swedish FrameNet++*

Lars Borin | University of Gothenburg

https://doi.org/10.1075/nlp.14.09bor



Pages 221–260 of
The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications
Edited by Dana Dannélls, Lars Borin and Karin Friberg Heppin
[**Natural Language Processing**, 14] 2021. xiv, 333 pp.

CHAPTER 9

Multiword expressions – a tough typological nut for Swedish FrameNet++[1]

Lars Borin
University of Gothenburg

Multiword expressions have attracted much attention in language technology over the last two decades or so, and in general linguistics, the interest in phraseology – which includes the linguistic study of multiword expressions – goes back much further. In our work on the multilingual components of Swedish FrameNet++, we have strived to adopt a typologically informed view on multiword expressions. This raises a number of theoretical and methodological questions, some of which are discussed in this chapter.

If you don't know where you're going, you might not get there.
Yogi Berra

1. Background

Multiword expressions (MWEs) have attracted much attention in language technology (LT) over the last two decades, at least since the publication of Sag et al. (2002). In particular, the international PARSEME initiative[2] has prompted a number of publications in recent years, addressing various aspects of representation and processing of MWEs in LT (see, e.g. Sailer & Markantonatou 2018; Markantonatou et al. 2018; Parmentier & Waszczuk 2019; Schulte im Walde & Smolka 2020), and also resulted in some valuable datasets (see Section 3.3.3).

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).

1. Linguistic examples in this chapter are glossed using the *Leipzig Glossing Rules* https://www.eva.mpg.de/lingua/resources/glossing-rules.php, with the following addition(s): HAB: habitual; VBZ: verbalizer.

2. https://typo.uni-konstanz.de/parseme/

https://doi.org/10.1075/nlp.14.09bor

This made it a very obvious desideratum to consider carefully how to best include MWEs in Swedish FrameNet++ (SweFN++). Our thoughts on exactly how this should be done were naturally informed by the fact that SweFN++ contains a massively multilingual component added as the result of two typologically and areally oriented linguistic projects (see Chapter 6 in this volume). Thus, we try to approach the question of how to describe MWEs in SweFN++ from a broad typological point of view – even if the concrete examples below come mainly from Swedish – and to consider what additional descriptive devices will be required in addition to those needed for lexical description of single-word expressions (SWEs).

However, the broad comparative approach characteristic of research in linguistic typology seems to have played a miniscule or non-existent role in both LT-oriented and linguistic work on MWEs. Comparative studies of MWEs in LT (or phraseology in linguistics) have generally been contrastive rather than typological in scope (van der Auwera 2012), i.e. they deal with (a convenience sample of) a few languages – typically only two – rather than with a systematic typological sample, which in the case of MWEs arguably should be a "variety sample", i.e. with one representative of every distinct genetic linguistic unit currently recognized (Bakker 2011), since we do not know the range of variation of this phenomenon. Taking the *Ethnologue* (Eberhard et al. 2021) as the basis for genetic classification of the world's languages, a minimal variety sample should contain ~120 languages, including all the language isolates recognized by the Ethnologue, such as Basque, Haida, Kusunda, Yukagir, etc.[3]

Linguistic typology is broadly concerned with uncovering the limits, distribution and interdependence of various linguistic phenomena in the languages of the world. It is a data-driven endeavor, relying on samples of many and diverse languages in order to cover the full breadth of linguistic diversity. Adopting a typologically informed view on MWEs raises a host of theoretical and methodological questions, which are the topic of Section 3. Even though SweFN++ is primarily an LT endeavor, this chapter is mainly about linguistic and lexicographical description of MWEs, and not about, e.g., how to find MWEs automatically in text or how to parse them. This is a related, but still distinct, and quite intricate set of problems (see, e.g. Parmentier & Waszczuk 2019; Constant et al. 2017).

By way of background to this discussion, the descriptive principles adopted at present for including MWEs in SweFN++ are presented in the next section.

---

3. If we instead use the more conservative classification adopted by the *Glottolog* (Hammarström et al. 2020) as our point of departure, we need to sample ~300 languages. In either case, far more than just a few languages are needed.

2. Multiword expressions in Swedish FrameNet++

A description of the treatment of MWEs in SweFN++ boils down to describing how they are handled in Saldo, the "pivot" resource of SweFN++. Saldo is described in much more detail in Chapter 3 in this volume (see also Borin, Forsberg, et al. 2013). Saldo is a full-sized modern Swedish lexical resource primarily intended for LT applications. It provides lexical-semantic, inflectional and compounding information on more than 147,000 entries. Saldo is an *onomasiological lexicon*, i.e. its entries designate *lexical senses*. Relevant in the present context, many entries (about 8,000 entries or ~6% of the total) are MWEs.

MWEs in Saldo are defined more or less as by Sag et al. (2002: 2), i.e. as lexicalized (or even conventionalized) expressions containing spaces in their written form according to the standard orthography of Swedish, i.e. the primary necessary criterion for MWE-hood is *orthographic*.

A good deal of thought has gone into integrating MWEs in Saldo in a way that is both practical and linguistically satisfactory. At the moment, we distinguish three different kinds of MWEs. These types are convenient to distinguish for (written) Swedish, and no claim is made as to their universality, nor that this list is exhaustive:

1. *Contiguous MWEs*; these correspond to the "fixed expressions" and "semifixed expressions" of Sag et al. (2002). The contiguity is on the level of lexemes, not, e.g., characters. Thus, the component lexemes may exhibit any combination of internal and external inflection. For example, the MWE *enarmad bandit* [one.armed bandit] 'slot machine' has the indefinite nominative plural *enarmade banditer*. However, the order of the constituent words is fixed and other sentence material (other words) never intervenes between the parts of the MWE.
2. *Noncontiguous MWEs*; these are, by and large, the "syntactically-flexible expressions" of Sag et al. (2002). In these, other sentence material may intervene, and the order of the parts may vary. Prototypical examples are particle verbs (see Example 1) and support verb constructions, i.e. constructions where a "semantically empty" verb is combined with a nominal (or adjectival) verb argument – often formally a direct object – which is the actual bearer of the predicate semantics, e.g. *draw a conclusion, take a walk, give a lecture, make an assumption*, etc.
3. *Constructions*; these are the kinds of phenomena that are studied in linguistics under the heading of *construction grammar* (Hoffmann & Trousdale 2013). MWEs are found among partially schematic constructions, i.e. syntactic fragments (or templates) with one or more slots for items specified as to, e.g., part of speech (in a dependency framework) or phrase type (in a constituency framework), and semantic type.

The first two kinds of MWEs are fully integrated descriptively in the Saldo morphology, and partly integrated with respect to morphological processing, while those falling under the third MWE category, the constructions, are left out of Saldo for the time being, as being the least "lexical", in terms of their formal behavior. However, a contributing project to SweFN++ has undertaken to build a Swedish *constructicon*, where a number of Swedish constructions are given a formalized linguistic description,[4] although not one immediately applicable in an LT context (Lyngfelt et al. 2018).

For the first two MWE types, we simply assume "word-like" – i.e. lexical – semantics: we treat them as *lexemes* in the sense used below in Section 3.1. You need not learn very many languages in order to observe that a single orthographic word in one language may correspond to a conventionalized orthographic MWE in some other language. The fact that such MWEs sometimes have compositional, non-MWE readings in addition to the conventionalized/lexicalized one is in principle no more of a theoretical problem than when a lexicalized compound also has a compositional reading in a language like German or Swedish (but it may of course present a very concrete practical problem for automated text processing). Cf. the Swedish compound *husbil* 'camper, trailer, RV', but also compositionally 'house car' (e.g., it could be used to refer to a builder's van with a drawing of a house on the side).

How often is a lexicalized MWE used with the alternative compositional reading? There is very little information available about this in the literature. A rare and welcome exception is the recent study by Savary et al. (2019), who investigate this for verbal MWEs in a corpus study of a small sample of languages, and find that literal readings account for approximately only 2% of the verbal MWE tokens in four out of the five languages investigated by them (Basque, German, Greek, and Polish), but for over 4% in the fifth language (Portuguese). In any case, the vast majority of instances carries the non-compositional meaning.

Even though compounds have been better studied than MWEs, this is not known about compounds either (at least I am not aware of any linguistic studies addressing this question), only that the compositional reading is possible as an alternative to the conventionalized or lexicalized meaning if all the component parts of the compound are also lexemes in the language. Intuitively, a compositional reading of a conventionalized compound normally has to be forced, and is typically construed as a pun, which indicates that this is not the normal state of affairs. If we assume that Swedish compounds are similar to MWEs in this respect, this intuition would be supported by the cited study by Savary et al. (2019). Note that even with a

4. https://spraakbanken.gu.se/karp/#?mode=konstruktikon&lang=eng

compositional reading, the semantic relationship between compound parts is underdetermined regardless of whether the compound is written as one orthographic word or separated,[5] which undoubtedly is a factor facilitating lexicalization.

The same semantic indeterminacy holds for derivational morphology, although to a lesser extent, since derivational affixes and processes tend to make a more specific semantic contribution to the resulting expression. Still, lexicalization is common here, too, so that e.g. the English denominal verb *knife* does not normally refer to any kind of use of a knife, but only to talk about stabbing. In the same way, the Finnish denominal/deadjectival noun-forming suffix *-sto/-stö* confers a general meaning of 'collective or collection', but in practice the derived words often have quite specific meanings: *kirjasto* (< *kirja* 'book') 'library', *vuoristo* (< *vuori* 'mountain') 'mountain range', *vähemmistö* (< *vähempi* 'lesser') 'minority', *miehistö* (< *mies* 'man') 'crew'.

In other words, there is plenty of scope for conventionalization/ lexicalization with all kinds of word-formation processes, including those yielding MWEs.

MWEs in Saldo are not described as having an internal syntactic structure, only an inflection table and a set of compounding forms. This is completely analogous to the treatment of structurally complex single-word items. We do not let the compound *husbil* inherit its formal characteristics from its last member *bil* 'car', but rather provide it with its own inflectional information, as if it were a simplex word. This is not to deny the value of such a description, which is what we expect to find in linguistic works on word-formation. In this regard we have simply opted to follow normal lexicographical practice, in not making the formal structure of complex words – compounds or derivations, and now also MWEs – explicit in the lexicon (Gantar et al. 2019: 139). In fact, a kind of conceptual paradox hides here, implicitly recognized by Haspelmath (2015: 297): If regular syntactic constructions are necessarily compositional – which seems eminently reasonable – MWEs are *automatically disqualified* from being analyzed as such constructions (NPs, VPs, PPs, serial verbs, etc.), simply by virtue of being non-compositional (except for trivially forming one-"word" NPs, VPs, etc.).[6]

---

5. When Swedish compounds are written as one word, there is also a segmentation problem, due to an orthographic rule prohibiting three identical consonants in a row. Thus, Swedish *glasskål* can be segmented in three different ways: *glas-skål* [glass-bowl], *glass-skål* [ice.cream-bowl], *glass-kål* [ice.cream-cabbage]. Of course, MWE analysis is also beset with similar ambiguity problems (Nasr et al. 2015).

6. But, of course, the possibly regular pathways by which fully compositional instances of regular patterns are lexicalized – and which regular patterns are amenable to such lexicalization – should be at least as interesting to language typology – which is the context of Haspelmath's remark – as frequently occurring grammaticalization pathways.

Consequently, we treat contiguous MWEs formally as "words with spaces", and subject to general morphology-like inflectional processes. We have yet to encounter some formal mechanism in such Swedish MWEs, which we would not also expect a general (inflectional) morphological processor to handle, in the sense that every such inflectional mechanism is attested as appearing word-internally (in SWEs) in some language ("internal" inflection, discontinuous dependencies among word components, multiple discontinuous exponence, coreference to word-internal components, and others which are attested as inflectional mechanisms among the languages of the world; see, e.g., Nida 1949; Jensen 1990). Thus, the formal behavior of Swedish MWEs does not warrant special treatment in this respect, as seen in a broad cross-linguistic perspective, even if in Swedish, some of the inflectional devices just mentioned are exclusive to MWEs.

With the noncontiguous MWEs, things become a bit more complex. The components of verbal MWEs (and sporadically MWEs from other parts of speech) can appear discontinuously in clauses. In theory, the intervening items can be arbitrarily long, but in practice they tend to be short, typically one to two words, as in (1) with the verbal MWE *göra sig till* 'posture; dissemble; playact', with both a reflexive pronoun (*sig*) and a particle (*till*):

(1) Swedish (swe) (Indo-European, Sweden, Finland; own knowledge)
Då **gjorde** hon **sig** verkligen **till**.
then **do.PST** she **3.REFL** really **to**
'Then she was really posturing.'

However, we still aspire to treat these kinds of verbal MWEs as lexical items rather than syntactic constructions: the description in Saldo is in terms of word semantics, and the formal treatment is one of "sequences with holes".[7] In part this decision has been motivated by the existence in languages of – formally not so different – mechanisms such as *incorporation* (Mithun 1984; Aikhenvald 2007) and *polysynthesis* (Fortescue et al. 2017), which are often considered to be lexical rather than syntactic in nature. Examples (2a)–(c) illustrate incorporation (as well as polysynthesis), while Example (4) shows polysynthesis without incorporation.

7. Note that we distinguish between verb particles and valency-bound prepositions. In Swedish, the former but not the latter carry primary stress. In this way, the (written) minimal pair illustrated by the expression *att hälsa på någon* [to greet on somebody] is not ambiguous in speech. The sense 'to visit somebody' (main stress on *på*; particle verb plus direct object) is clearly distinguished from 'to greet somebody' (main stress on *hälsa*; simplex verb plus preposition-phrase complement).

(2) Chukchi (ckt) (Chukotko-Kamchatkan, Russia; Skorik 1961: 101–103)
   a. tə-takečγə-pelja-rkən
      1SG-meat-leave-1SG.IPFV
      'I leave meat'
   b. tə-pəlvəntə-kopra-ntəvatə-rkən
      1SG-metal-net-set.out-1SG.IPFV
      'I set out a metal net'
   c. tə-vel-ənnə-tke-rkən
      1SG-rotten-fish-smell-1SG.IPFV
      'I smell of rotten fish'

A consequence of the foregoing is that we assume the same set of parts of speech (POS) for MWEs as for single-word lexical entries. Thus, *svart hål* 'black hole' is a (multiword) noun, *skriva ut* [write out] 'prescribe (medicine, etc.); discharge (from hospital, etc.); print' a (multiword) verb, and *med andan i halsen* [with breath.SG.DEF in throat.SG.DEF] 'breathlessly' a (multiword) adverb. Formally, the POS label of MWEs is formed by suffixing an "m" to the corresponding SWE POS label: "nnm" is a multiword noun, etc.

The tricky cases include full clauses or sentences, e.g., proverbs such as *bränt barn skyr elden* [burnt child shuns fire.SG.DEF] 'once burnt, twice shy'. Rather than introducing a "clause" part of speech or treating these as zero-argument verbs (which otherwise do not occur in Swedish), they are classified in Saldo as multiword interjections (inm); like interjections (and vocatives), they are not normally properly parts of the clauses they appear in, although they can appear in nominal slots in *de dicto* usages, again like interjections.

Even if this chapter is primarily about the *description* of MWEs in SweFN++, and not about their *processing* in LT systems, we would like to note that neither fully compositional compounds nor, e.g., fully compositional particle verbs, adjective–noun combinations or prepositional phrases, should be listed in the lexicon. Since both many compounds and many particle verbs in texts originate in regular constructions, and since many of them also have conventionalized or lexicalized senses, processing components which use the lexical resource should also include the facilities for on-the-fly compositional analysis of both MWEs and (SWE) compounds. In other words, the fact that our lexicon contains an entry *husbil* 'camper, RV' should not exclude the regular compound analysis *hus-bil* [house-car] from being made, just as the listing of *ta upp* 'bring up/raise (an issue)' should not prevent the regular compositional alternative analysis [take up] 'pick up (e.g. an object from the floor/ground)'.

3. MWEs from a typological perspective: A first cut

In the LT literature as well as in linguistic works on phraseology we encounter a number of general statements – claims and hypotheses – about MWEs, which have been formulated on the basis of data from a single or a few languages. The main question for the present chapter is to assess such statements against a broader cross-linguistic background, with a view to couch the treatment of MWEs in SweFN++ in terms enabling broad typological comparison.

3.1 The "words" of MWEs

Baldwin & Kim (2010: 269) propose a "formal definition" of MWEs: "lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity". Paradoxically, as Baldwin and Kim 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.[8] Baldwin & Kim (2010) mention German compounds here, but their definition would arguably apply equally to, e.g., noun incorporation as found in many languages all over the world (Mithun 1984); see Examples (2a)–(c). It also logically allows for MWEs being made up of other MWEs, which is easier to accept.

In Baldwin and Kim's 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' (Aikhenvald & Dixon 2002; see also Haspelmath 2011b; Bickel & Zúñiga 2017). Although a quite central concern to LT, the 'orthographic word' is not generally accorded much weight or even explicitly recognized in typologically oriented linguistics, since (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, and not even be consistent.

For instance, there may be both a single-word and a multi-word spelling recognized as representing the same lexical item, e.g., Swedish *idag* ~ *i dag* 'today'. The same thing can be observed for English compounds, e.g.: *greenhouse* ~ *green house* ~ *green-house*. In Bantu linguistics, a distinction is sometimes made between "conjunctive" and "disjunctive" orthographies, where these terms refer to the differing practices in different Bantu languages of writing cognate subject and object indexing

8. For instance, Sag et al. (2002: 2) talk explicitly about "idiosyncratic interpretations *that cross word boundaries (or spaces)*" (emphasis added).

morphs and other elements signalling the syntactic role of verbs as either prefixes of verbs (conjunctive) or as separate words (disjunctive) (Taljard & Bosch 2006).

A language may also exhibit orthographic variation dependent on lexicogrammatical factors. In Swedish and Finnish, we find both multiword and single-word forms among derivationally related lexical items based on verbs. Swedish particle verbs normally appear with the components separated, showing the order verb–particle, and other clause elements may intervene between the verb and the particle, e.g., subjects and sentence adverbials; see Example (1). The same items form past participles where the particle is joined to the verb as a prefix: *tappa bort* [lose away] ~ *borttappad* [away.lost] 'lose (an object)' ~ 'lost'.[9] In many Finnish compounds – which are written as one word according to Finnish orthography – where the head is a participle, or a deverbal actor or action noun, the components can be written separately as well, without any other change in form (Hakulinen et al. 2004: § 418), e.g.: *tarjoilu pöytiin* [serve.NMLZ.SG.NOM table.PL.ILLATIVE] ~ *pöytiintarjoilu* ~ *pöytiin tarjoilu* 'table service'.

In older, historical varieties of Swedish, compounds have frequently been written with components separated, and are often formally indistinguishable from regular possessive constructions, i.e. the first part will have its expected genitive form according to its declension (see Chapter 4 in this volume). Even today, there is a tendency to write Swedish compound components separated, especially in social-media genres and on non-official signage, a tendency often attributed to influence from English.

Here we are dealing with an issue at whose heart lies a conceptual or methodological conundrum, namely the difficulty of providing a definition of "word" that will work for all languages. Haspelmath (2011b) 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". Getting ahead of ourselves a bit, we could say that the definition proposed by Baldwin & Kim (2010) actually comes close to what would be needed for a serious typological study of MWEs (see Section 4 below), exactly because it downplays the importance of the (orthographic) word for delimiting the phenomenon of interest.

9. Analogous cases are found in German and Hungarian, and Savary et al. (2018) consistently annotate such items as MWEs even when written as single words, according "lexeme unity" priority over orthography, as it were.

3.2 The "lexemes" of MWEs

Even if we decide to understand the "W" of "MWE" as 'lexeme' rather than, e.g., 'text word', there still remain some conceptual and practical issues to resolve.

How are we to think about expressions containing only one open-class item, e.g., *off the hook, on edge*, or *at least*. The Swedish (orthographic) MWE adverb *i klistret* [in glue.SG.DEF] 'in a pickle' could be translated in its idiomatic reading into Finnish as *liemessä* [broth.SG.INESSIVE], i.e. into a single-word expression, a case-inflected noun.

At issue here is that the term "lexical" is used in (at least) two quite different ways in the linguistic literature, a relevant and even important difference in our context. It may mean roughly 'being listed in a (conventional) lexicon; pertaining to vocabulary', but also 'belonging to the content words'. The latter usage can be seen, e.g., in the characterization of the Universal Dependencies (UD) formalism as one which "strongly prefers lexical heads" (Silveira & Manning 2015: 310). The matter is complicated by the fact that "grammatical" can function as the opposite – or at least complementary term – in both cases (although the UD literature prefers to set "lexical" against "functional"). This means that at least potentially the requirement that an MWE contain more than one "lexical item" could be understood as excluding expressions with only one content item. This does not seem to be the case in practice, however.

On the semantic side, the main interest of researchers who have studied MWEs seems to have been focused on degree of compositionality. MWEs are distributed along a continuum, from those at one end that show only a collocational preference in the choice of synonymous words, over partly interpretable MWEs, to full idioms at the other end of the scale. Contrary to this and not surprising, lexicographers recognize only a binary opposition: a complex linguistic expression is either to be listed in the lexicon or not.

What is considered compositional and not is of course dependent on one's view on word senses. In LT work it has long been recognized that too fine-grained word sense inventories – such as the 59 senses of the verb *break* in Princeton WordNet (PWN) – are difficult to distinguish reliably to machines and people alike, with the possible exception of highly trained lexicographers (Kilgarriff 1997; Hanks 2000). However, PWN simply follows (Anglo-Saxon) lexicographic tradition here; at https://www.dictionary.com/browse/break (based on the *Random House Dictionary*) 69 senses are given for the verb *break*.

When it comes to degree of colexification[10] in lexical description, lexicographical traditions are situated along a "lumping"–"splitting" axis. Ultimately,

10. *Colexification* was introduced in the context of lexical typology by François (2008) as a neutral superordinate term for polysemy and homonymy.

this parameter is intimately tied up with how we conceptualize language and the linguistic knowledge involved in understanding and producing language.

The traditional English-language lexicography exemplified by PWN leans toward the "splitting" camp, whereas the Swedish tradition underlying SweFN++ is more of the "lumping" kind, which concretely manifests itself in there being close to an order of magnitude less colexification in Saldo, as compared to PWN.

We should expect a "splitting" lexicographical tradition to recognize fewer MWEs than a "lumping" one.[11] This is because the former tradition seems to be predicated on a strict notion of compositionality, in the extreme cases including "lexical items" defined through idiosyncratic decomposition of expressions which "are decomposable but coerce their parts into taking semantics unavailable outside the [multiword expression]" (Baldwin et al. 2003: 89; see also Nunberg et al. 1994).[12] Even in the ordinary case, strict compositionality means, roughly, that there is no scope for rich general rules of inference in interpreting linguistic expressions; rather, words should carry as much as possible of their interpretation in each specific context with them, which potentially leads to as many meanings as there are distinct contexts, and consequently, to the postulation of fewer MWEs. This stance stands in contrast to one which posits more general "meaning potentials" (Hanks 2000, 2002, 2013) or even overlapping senses (Erk 2010) for lexical units, which would instead rely on a sophisticated and information-contributing interpretation procedure on the part of the language user, plus a larger share of "prefabs", i.e. MWEs.

While it may not be too hard to formulate criteria that let us discover (at least some kinds of) formal idiomaticity, the situation is much more unclear when it comes to lexicalization or conventionalization of complex linguistic expressions. As we have just seen, the converse notion of compositionality is heavily dependent on theoretical and methodological assumptions ("prejudices"). We would be greatly helped in our efforts by knowing, e.g., the limits – if there are any – of lexicalization, both in terms of the kinds of concepts that will be amenable to lexical coding and in terms of the formal constructions that can become conventionalized as lexical items. These are in principle research questions for *lexical typology* (Koptjevskaja-Tamm et al. 2007) and *semantic typology* (Evans 2011). So

---

11. Although logically we would of course also expect a higher degree of colexification in those MWEs that a "splitting" tradition would recognize as such!

12. E.g.: "More specifically, human beings have a natural tendency to define the context (without admitting, even to themselves, that they are doing so), rather than focusing on the particular contribution of the word to the contexts in which it occurs. An extreme example, mercifully cancelled before publication, was a proposed definition of ***throw*** as 'to behave in a wild and uncontrolled manner'. When challenged for evidence, the lexicographer who wrote it pointed to the expressions '*throw* a fit' and '*throw* a wobbly'." (Hanks 2002: 159; original emphasis)

far, intuitions have been on the pessimistic side as to the feasibility of formulating broad cross-linguistic generalizations in this area; see Chapter 6 in this volume.

However, this must ultimately be an empirical question, and I believe that LT can provide excellent tools and methods for looking for answers to it.

3.3 How frequent are multiword expressions in language?

The question of how frequent MWEs are in language can refer to their share of the vocabulary or their text frequency.[13] In both cases, the answer is relevant to LT.

3.3.1 *MWEs in the lexicon*

Can we make any claims about the preponderance of MWEs in specific languages or even propose a typological classification of languages based on this?

So far, there are no empirically well-founded such figures for any language. 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*. However, it is also supported by the corresponding PWN statistics, where MWEs make up a considerable share of the entries, from approximately one third to over 40%, depending on how they are counted; see Section 3.3.2 below.

The MWEs in Jackendoff's dataset are made up of multiple orthographic (English) words. Given that compounds make up a sizeable share of these MWEs (a third of all MWEs in the dataset, and about half if names, titles and quotations are excluded), then the corresponding estimate for languages such as Swedish, Finnish or German, where compounds are written as one orthographic word, should be that they have – *ceteris paribus* – on the order of at least a third as many MWEs as SWEs, if names, titles and quotations are excluded, i.e. about 25% of the lexicon should be made up by MWEs (as opposed to 50% in English), and another 25% should consist of compounds.

Since Saldo follows the traditional lexicographic principle referred to above of not providing information about the internal structure of formally complex lexical items, it does not contain explicit information about compounds, but it is not difficult to extract Saldo entries like

örnnäsa : näsa + *böjd* 'eagle nose' : 'nose' + 'hooked'

13. "Text frequency" is intended to also include frequency in (unwritten) speech.

i.e. entries with a primary descriptor whose POS is the same as the entry and whose lemma is a suffix of the entry's lemma.[14] This does not capture very idiomatic cases, such as

droppnäsa : avleda + byggnad / vatten
[drip.nose] 'drip edge, drip strip' : 'divert' + 'building' / 'water'
skärgård : ö[2] [skerry.yard] 'archipelago' : 'island'

where *droppnäsa* has *näsa* 'nose' as second element, but the word does not denote a kind of nose (except perhaps metaphorically), and *skärgård* like a number of other nouns ending in *-gård* reflects an obsolete meaning of this element.

This yields 28,580 hits in Saldo, out of which 25,614 are nouns and 1,147 (almost exclusively prefixed, not compounded) verbs.

If we include entries where we also look one level further up in the Saldo hierarchy, at the primary descriptor's primary descriptor, we will find entry combinations such as (literal translations in square brackets)

jordskredsseger : valseger + enorm
'landslide victory' : 'election victory' + 'enormous'
valseger : seger + val[4] 'election victory' : 'victory' + [choice[4]] 'election'

This is a common pattern in Swedish nominal compounding, where a "logically" expected three-member compound comes out having only two members, which means that the meaning of these compounds cannot be compositionally derived. Some other examples:

grythund : jakthund + gryt 'burrowing dog' : 'hunting dog' + 'burrow (n)'
stegbil : brandbil + stege
[ladder.car] 'ladder truck, fire truck' : [fire car] 'fire engine' + 'ladder'
temanummer : tidskriftsnummer + tema
[theme.number] 'thematic issue' : [journal number] 'journal issue' + 'theme'

There are almost 2,500 such entries in Saldo, and among these the nouns dominate even more, with 2,348 nouns against 59 verbs.

---

14. Each entry in Saldo – a Swedish word sense – is characterized by one or more so-called *descriptors*, minimally a *primary descriptor* which is a both more central and semantically maximally close neighbor of the entry. In practice, the primary descriptor will often be a synonym or hyperonym of the entry. The optional secondary descriptor(s) typically add(s) some disambiguating information helping a reader to pinpoint the intended word sense. See Chapter 3 in this volume for a more detailed description of the structure of Saldo.

Even if we do not find all the compounds in Saldo in this way, these results give a fair idea of the magnitude of the compound component in Saldo. All in all, at least a fifth of all entries (and a third of the nouns) in Saldo are nominal compounds, a proportion not far off from the 25% guesstimated above on the basis of Jackendoff's (1997) calculation of MWE incidence in English.

Saldo contains 5–6% (orthographic) MWEs – depending on how they are counted; see Tables 1 and 2 and the discussion below in Section 3.3.2 – which indicates either that many MWEs are still missing from our lexicon or that Swedish and English are significantly different in this regard.[15]

English is in no way extreme in this regard. There are languages such as Kalam, with no more than 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):

(3) Kalam (kmh) (Trans-New Guinea, New Guinea; Pawley 1993: 95)
b ak am mon p-wk d ap ay-a-k
man that go wood hit-break get come put-3SG-PST
'The man fetched some firewood.'

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), cf. (4).[16]

(4) Inuktitut (ike) (Eskimo-Aleut, Canada; Dorais 2017: 135)
sinnatuuma-ju-ujaa-raalut-tu-ujaa-nirar-
to.dream-INTR.PTCP-look.like-much-INTR.PTCP-look.like-say.that-
-ta-u-qatta-lau-runnai-nira-laur-tu=ugaluaq
-PASS.PTCP-be-DUR-PST-not.anymore.say.that-PST-3SG.ind=however
'However, he said that it was not unusual anymore for him to be said to look like somebody who looks a lot like one who is dreaming'

These facts indicate that languages may differ as to the number of MWEs present in their lexicon, even with different definitions of MWEs. This prompts the following concrete questions:

---

15. The absolute numbers in Tables 1 and 2 are a snapshot taken in January 2021 during the writing of this chapter. Saldo is constantly growing, so these numbers change. Proportions (percentages) will be more stable.

16. This is one word: *sinnatuumajuujaaraaluttuujaanirartauqattalaurunnainiralaurtuugaluaq*, which does not fit on one line in Example (4).

1. Are there languages without MWEs? The general view in the literature seems to be that MWEs are universally present in languages, but this has not been systematically investigated.[17]
2. What is the minimum and maximum share of MWEs in the lexicon of any language? In texts? If there is cross-linguistic variation in this respect, is this variation correlated with other typological features, language-internal or language-external (sociolinguistic or population variables such as the size of the language community or its proportion of L2 speakers)?
3. How diachronically stable are MWEs and MWE types?

With respect to the last question, note that we cannot expect to reconstruct a past language stage without MWEs, if we have answered the first question in the negative for existing languages, nor more or less MWEs – on average – in historical language stages than in contemporary languages. Historical linguistics swears – indeed, must swear – by the *uniformitarian principle*:

> […] Nothing (no event, sequence of events, conjunction of properties, 'general law') was ever the case only in the past.
>
> I.e., the general principles that govern the world in the present hold for the past as well. Without this control, nothing can stop us from reconstructing anything at all
> (Lass 1978: 276f)

It would also be interesting to investigate the conditions under which MWEs can be borrowed, as well as the extent to which particular MWE-forming mechanisms will spread from language to language in language contact situations.[18] E.g., Swedish is quite open to importing English particle verbs by borrowing the verb part (if necessary) and combining it with the corresponding cognate Swedish particle: English *sign up* becomes Swedish *signa upp* (or *sajna upp*), where the verb is borrowed and the particle a cognate substitution.

---

17. As a parallel, there seem to be languages without compounds, and among those that do have them, their incidence varies greatly (Bauer 2009).

18. Thus, Aikhenvald (2006: 52) states that "[v]erb serialization as a grammatical mechanism tends to diffuse", and according to Ciancaglini (2011), support verb constructions ("periphrastic verbs") have been borrowed into a number of Semitic languages and possibly also Turkish from Iranian languages.

Table 1. POS distribution among Saldo SWE and MWE entries

POS	SWE	MWE	MWE %
Adjective	33,946	402	1.17
Adverb	1,139	1,872	62.17
Conjunction	12	5	21.41
Interjection	260	265	50.48
Noun	83,689	673	0.80
Numeral	104	1	0.95
Preposition	179	95	34.67
Pronoun	93	87	48.33
Proper noun	5,318	423	7.37
Subjunction	46	27	36.99
Verb	7,925	4,501	36.22
Total	**132,711**	**8,351**	**5.92**

3.3.2 *Which kinds of lexical units should we count?*

Saldo has two kinds of lexical units. Word senses form the primary entries of Saldo and lemgrams identify the formal (inflectional and compounding) behavior of the text words which express the word senses. Both are identified by a kind of "lemma", a conventional representation of the lexical unit in question, which contains a (version of a) traditional citation form in order to make it conveniently human-readable (see Chapter 3 in this volume). Consequently, in both cases we will find the expected distinction into SWEs and MWEs reflected in the orthography of the identifier itself.

The proportion of MWEs in the lexicon becomes slightly different, depending on whether we take the word senses or the lemgrams as the units under investigation: while the proportion of MWEs among word senses is 5.18%, its share of the lemgrams is 5.92% (see Table 1) for their POS distribution.

Similarly, in the downloadable version of PWN 3.0[19] the share of word senses with a MWE lemma is ~33%, and the MWE proportion of the distinct lemmas is almost 44%, while MWEs make up about 41% of the unique lemma–POS combinations present in PWN.

I believe that the latter way of counting – i.e. Saldo lemgrams or PWN lemma–POS combinations – is the most reasonable, since multiwordhood is primarily a characteristic of the *form* of linguistic expressions, where form is naturally understood to include part of speech membership. Semantics does come into the picture, but the form criterion is primary. If MWEs are not systematically different from

19. https://wordnet.princeton.edu/download

SWEs in colexification potential – which must surely be the null hypothesis – the ratio between the two ways of calculating should stay stable.

3.3.3 *MWEs in texts*

Jackendoff's statement cited in the previous section concerns the lexicon (of English), as do the statistics for Swedish shown in Table 1. We would also like to have some information about the *text frequency* of MWEs. This information would be valuable in its own right, and also useful for building LT systems. As one of the few concrete empirically determined figures found in the literature, Nivre & Nilsson (2004: 41f) report approximately 2 MWEs per 100 words of running Swedish text.

A considerably more ambitious undertaking is described by Savary et al. (2018). The PARSEME corpus of verbal MWEs provides text statistics for verbal MWEs in 18 languages. An approximate estimation based on the figures provided by them (Savary et al. 2018: 118ff) gives a ballpark figure of 1–2% verbal MWEs in most of the 18 languages represented in their corpus, with some notable exceptions, discussed in Section 3.3.4 below.

How do we reconcile these low numbers with the notion that MWEs make up a substantial portion of the lexicon, supposedly at least one fourth in Swedish and at least half in English? One possible reason for this seeming discrepancy could be formulated along the following lines. Word length and text frequency correlate, so that higher-frequency words tend to be shorter: "The magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences." (Zipf 1935: 25). If MWEs are "words" in this sense, they will by necessity on the whole be longer than single words, and we would in fact expect the MWEs not to exhibit high text frequency, even if they are pervasive in the lexicon.

We can use Saldo in order to partly test this explanation. The organization of Saldo is hierarchical, defined by a lexical-semantic relation which places more central word senses higher up in the hierarchy, closer to the root of the tree. High text frequency of a lexical item is supposed to correlate not only with length, but also with centrality, or "coreness", of this item, so that lexemes with high text frequency are expected both (1) to be found high up in the Saldo hierarchy and (2) to be short (see Chapter 3 in this volume). Table 2 shows the proportion of MWEs for each level of the Saldo hierarchy (1 being the highest, most central level, and 16 the lowest).

In Table 2 we see that the generalization just stated seems to hold only partly. The MWE share of the vocabulary reaches a maximum at depth 2, although the average entry length (all entries, not only MWEs) increases monotonically down through level 10, as expected. For comparison, the statistics for compound nouns calculated as described above in Section 3.3.1 are also given in the table (although also included in the SWE counts).

Table 2. SWE and MWE word senses per Saldo level
(CNs = compound nouns [included in SWE counts])

Saldo level	Number of				Percent		Entry length (chars)			
	Entries	SWEs	(CNs)	MWEs	(CNs)	MWEs	Mean SWEs	Mean MWEs	Mean all	Mdn all
01	41	41	0	0	0.00	0.00	3.63	0.00	3.63	4
02	970	793	66	177	6.80	18.25	6.81	10.90	7.56	7
03	9,064	8,302	674	762	7.44	8.41	7.55	11.65	7.90	7
04	18,295	16,924	3,811	1,371	20.83	7.49	9.16	11.73	9.36	9
05	30,060	28,181	6,740	1,879	22.42	6.25	9.55	11.86	9.69	9
06	31,838	30,254	6,806	1,584	21.38	4.98	9.93	12.04	10.03	10
07	24,575	23,541	4,673	1,034	19.02	4.21	10.15	12.18	10.24	10
08	15,801	15,202	2,583	599	16.35	3.79	10.27	12.33	10.35	10
09	9,109	8,820	1,425	289	15.64	3.17	10.35	12.51	10.42	10
10	4,405	4,266	718	139	16.30	3.16	10.50	12.34	10.56	10
11	1,833	1,773	302	60	16.48	3.27	10.33	12.05	10.39	10
12	673	661	109	12	16.20	1.78	10.48	13.25	10.53	10
13	233	232	24	1	10.30	0.43	10.00	15.00	10.02	9
14	53	53	2	0	3.77	0.00	9.89	0.00	9.89	9
15	13	13	0	0	0.00	0.00	10.38	0.00	10.38	11
16	2	2	0	0	0.00	0.00	14.50	0.00	14.50	14
all	146,965	139,058	27,933	7,907	19.01	5.38	9.72	11.95	9.84	9

On the other hand, for a number of reasons – both conceptual and empirical – we have very little knowledge of the true text distribution of MWEs in any language.[20] First, annotated corpora tend to lack the relevant information, and, second, automatic annotation tools do not provide it either. This would then be a concrete question for empirical cross-linguistic research: What is the range of attested text distributions of MWEs across the languages of the world?[21]

---

20. In fact, this is a potential confounding hidden variable for psycholinguistic experiments attempting to distinguish between degrees of compositionality in MWEs on the basis of reaction times, such as those reported by Gibbs et al. (1989).

21. There seems to be at least some relevant information available in the literature, e.g.:

> Surveying the languages described in this volume, we find that the following approximate percentages of textual clauses include [a serial verb construction]:
>
> – more than 70 per cent: Tariana
> – between 50 per cent and 70 per cent: Ewe, Eastern Kayah-Li, Dumo
> – between 20 per cent and 50 per cent: Goemai, Thai, Tetun Dili, Olutec, Cantonese
> – between 5 per cent and 20 per cent: Mwotlap, Toqabaqita, Lakota
> – less than 1 per cent: Khwe (Dixon 2006: 338)

3.3.4 *MWEs and parts of speech*

Baldwin & Kim (2010: 274) state that "[n]ominal MWEs are one of the most common MWE types, in terms of token frequency, type frequency, and their occurrence in the world's languages". This may well be true, but the references given in support of this statement (Tanaka & Baldwin 2003; Lieber & Štekauer 2009) do not actually provide this information, as far as I can see. Nor does the detailed overview by Aikhenvald (2007) clarify this issue, indicating that this may not be known, and consequently that further research is called for.

Table 1 presents the MWE statistics for some parts of speech in Saldo, where we see that nouns end up in third place, after verbs and adverbs (mainly prepositional phrases). The Saldo statistics are directly relevant to the third part of the claim ("occurrence in the world's languages"). However, if we add the estimation of compound nouns in Saldo made above in Section – formally (orthographic) SWEs in Swedish, but MWEs according to Baldwin & Kim's (2010) definition – nouns will come out on top by a comfortable margin in absolute numbers, although not if we look at how large a share of a part of speech is made up of MWEs. Thus, Baldwin & Kim's (2010) claim finds support in these numbers.

Nivre & Nilsson (2004: 42) provide some statistics illuminating the claim about token (text) frequency: adverb MWEs hold the top position in their material, showing almost twice as many occurrences as the MWE proper nouns. However, in their corpus, no MWE nouns have been annotated. Nor are any verbal MWEs reported by them, indicating that the MWE definition used was quite different to the compilers of the corpus compared to what we find in the current literature. In the Swedish part of the multilingual corpus described by Savary et al. (2018) the proportion of verbal MWEs is reported as 1%, which would give an overall MWE share of approximately 3% in Swedish texts, not counting MWE nouns.

In any case, the discrepancy between the claim and the statistics just cited reveals an obvious need for more data to be gathered and analyzed. Unfortunately, when it comes to MWEs, linguistic typological research has tended to focus on complex predicates, i.e. verbs (see Section 3.5 below), and very little work has been done on other parts of speech from a broad typological perspective. Notably, adverbs form the second-largest MWE category in Saldo in absolute numbers and the largest if we see to the MWE proportion of a particular part of speech. We also find many closed-class items among the MWEs: pronouns, adpositions, subjunctions, etc. When MWEs are considered, closed parts of speech become, if not open, at least less closed, as it were. In this connection we may note that lexical change in such cases may proceed via an intermediate MWE stage, as when Old French *ne* is expanded to Modern Standard French *ne … pas* [not … step(n)] and subsequently reduced to (colloquial) *pas*, all meaning 'not' (clausal

negator). This indicates that the role of MWEs in grammaticalization processes warrants further study.[22]

Even though the main focus of his investigation is on verbal MWEs, Pawley (1993: 99–101) mentions in passing that Kalam exhibits some nominal compounding and also makes extensive use of full clauses as nouns:

(5) Kalam (kmh) (Trans-New Guinea, New Guinea; Pawley 1993: 100)
bynb penpen ña-p-ay
people reciprocally shoot-HAB-3PL
'enemies' (i.e., 'people (one) fights with/used to fight with')

Verb-poor languages have been attested from various parts of the world (Kalam is far from unique in this regard), but as far as I know, noun-poor languages have not been reported (see Section 3.3.5). Given that verbs, nouns, and interjections are the universal SWE parts of speech (sometimes complemented by a closed class of "particles", which sometimes also includes the interjections), what conclusions can we draw about MWEs? As we saw above, even if Kalam has a very small number of SWE verbs, available core predicates are in fact numerous and expressed as MWE serial verbs. From Pawley's (1993) description it is even likely that the number of verb lexemes available to the languages users – mostly MWEs – is at least on a par with those of a language such as English. On a more general note, should we expect different distributions over parts of speech for MWEs in languages based on their SWE distributions?

3.3.5 *Towards a typological generalization?*

A most interesting study, highly relevant to one of the central questions posed earlier – in what way (if any) variation in MWE occurrence is correlated with other typological features – is that by Polinsky (2012) and Polinsky & Magyar (2020), where verb and noun lemma counts from corpora and dictionaries are used in order to investigate the relationship between verb-to-noun ratio (VNR)[23] and headedness in language. Their results show a clear correlation, in that head-final (HF) languages tend to have low VNR and head-initial (HI) languages have high VNR, while SVO (subject-verb-object) languages show a more varied picture. Relevant to our aims here, they also discuss why low VNR would indicate a significant use of support verb constructions (SVCs) in a language. Polinsky & Magyar's (2020) investigation has many points of correspondence with the work conducted over the years in the

22. It has been suggested that the rich array of verb-forming suffixes in e.g. the Wakashan languages are grammaticalized full (support) verbs (Mithun 1984: 887ff; See also Example 7).

23. Calculated as the number of verb lemmas divided by number of noun lemmas in a corpus or dictionary.

SweFN++ project, and raises many of the same theoretical and methodological issues that we have been grappling with in our work. I now turn to a discussion of some of the central such points.

3.3.5.1 *Data sources and data commensurability*

For their main study Polinsky & Magyar (2020) draw on corpora (Universal Dependencies – UD – corpora and various "national" corpora) and dictionaries for 35 languages. The sample is subject to the usual restrictions regarding availability of data, etc., and taking this into account, it appears to reflect an acceptable degree of genealogical and geographical diversity, given the aims of the study.

Neither dataset versions nor the procedure used for calculating POS statistics are specified. This makes replication of results difficult, a matter to which I return below.

For two languages (English and Russian) Polinsky & Magyar (2020) show that VNR calculated from corpora and from dictionaries come out more or less the same, at least if the corpora and dictionaries are large enough.[24] It would be good to be able to investigate this over a wider range of languages, simply to ensure that this is not dependent on some typological feature. Given the aim of Polinsky & Magyar's (2020) investigation, the expectation would have been that at least one instance of each language type should be checked in the same way, i.e. additionally at least one HF and one HI language.

Incidentally, some information can now be added to that provided by Polinsky & Magyar (2020). They report a VNR of 0.15 for Hungarian, calculated from the Hungarian National Corpus (although no details of the calculation are provided), which is quite in line with their hypothesis about the relationship between VNR and headedness. However, this figure is quite surprising, since Hungarian like several of its Finno-Ugric sister languages has a rich derivational morphological apparatus, including several very productive verb-forming suffixes, both for forming verbs from other parts of speech and for forming verbs from verbs, also indirectly, e.g. via an intermediate deverbal noun. Using a freely available Hungarian dictionary,[25] we find a VNR of approximately 0.39, more in line with what is intuitively expected.

---

24. Intriguingly, they also remark in passing that the percentage of nouns remains fairly stable over all investigated corpora and dictionaries. Since VNR varies considerably, this then raises the question: what compensates for the difference in verb percentages, if not the nouns?

25. Czuczor & Fogarasi (1862), downloaded from http://osnyelv.hu/czuczor/ on 2021-01-04. This dictionary contains around 100,000 entries. Unfortunately, I was not able to locate any more recent openly available dictionary on the internet. The downloaded html files (one for each letter of the Hungarian alphabet) were merged into one long file, and POS counts were found with command-line expressions like (in this case for counting nouns)

grep -A 1 'color=#ffffff' au8.txt | grep ' fn\. ' | wc -l ("fn." = "főnév" 'noun')

This made me attempt to replicate the results from the article, and using the downloadable frequency data for the Hungarian National Corpus[26] I counted verb and noun lemmas in this dataset,[27] resulting in 83,598 verbs and 431,196 nouns, giving a VNR of 0.19, not much higher than the VNR of 0.15 reported by Polinsky & Magyar (2020). HF Hungarian thus does not show such a close agreement in VNR between corpus and dictionary as Russian and English.[28] More research is obviously needed in this area.

Another remark by Polinsky & Magyar (2020) that I wished to follow up on was in regard to the unexpectedly large difference in VNR between Polish and Russian, in both cases based on the UD corpora. My attempt to reproduce the results failed, however. I downloaded the Polish and Russian UD training sets and calculated verb and noun statistics and VNR from them.[29] This yielded a VNR for Polish of 4,438/10,180 = 0.44, and for Russian of 7,192/14,551 = 0.49, which are much closer than Polinsky & Magyar's (2020) results, and the latter also have a higher VNR for Polish (0.56) than Russian (0.47), i.e. the opposite of what I found. Since Polinsky & Magyar (2020) do not provide information about which version of the UD corpora they have used, nor about how the POS information has been extracted, these results cannot be directly compared. Rather, they underscore that, again, there is much more work to be done in this area.[30]

---

26. Reachable via http://corpus.nytud.hu/mnsz/index_eng.html (accessed on 2021-01-05)

27. With a simple Unix command-line count:

```
cut -f 3,4 hnc-1.3-wordfreq.txt | grep -a '\tV$' | sort | uniq | wc -l
cut -f 3,4 hnc-1.3-wordfreq.txt | grep -a '\tN$' | sort | uniq | wc -l
```

28. It is probable that the lemma counts are inflated in the corpus, due to out-of-vocabulary (inflected) forms being listed as lemmas. I do not know if there is any systematic POS bias in this process, however.

29. https://github.com/UniversalDependencies/UD_Polish-PDB/raw/master/pl_pdb-ud-train.conllu ("Latest commit ce3e454 on 7 Sep 2020")
https://github.com/UniversalDependencies/UD_Russian-SynTagRus/raw/master/ru_syntagrus-ud-train.conllu ("Latest commit deca643 on 2 May 2019")
(both downloaded on 2021-01-05) In both cases the POS counts were obtained by executing the Unix shell commands

```
grep '^[0–9]' <CONLLU file> | cut -f 3,4 | grep '\tNOUN' | sort | uniq | wc -l
grep '^[0–9]' <CONLLU file> | cut -f 3,4 | grep '\tVERB' | sort | uniq | wc -l
```

Russian also has 2 instances of AUX and 7,049 instances of PROPN, which have not been counted. Polish also has 10 instances of AUX and 4,827 instances of PROPN, which have not been counted.

30. Also, when working with relatively large datasets in many languages and formats, it is very easy to slip up. A single misplaced delimiter or indentation in a processing script can have large consequences for the end result. This makes it even more imperative to both provide sufficient

One issue that Polinsky & Magyar (2020) mention in passing, but do not follow up, is that of corpus size. It is well-known that relative word frequencies are non-linearly related to corpus size (Baayen 2001). What should be ascertained is whether POS proportions remain the same regardless of corpus size. If the share of verbs grows slower than that of nouns with increasing corpus size, this will impact the results if not compensated for, e.g. by averaging over equal-sized sampling frames when comparing corpora of different sizes. This potential confound also applies to dictionary data, since we cannot *a priori* expect the growth rate for different parts of speech to be the same as the lexicon grows in size in the way that lexical resources are normally extended, i.e. from "core vocabulary" towards increasingly rare and specialized items. This is clearly an empirical matter, and I conducted a small initial experiment in order to test this. Two digitized Sanskrit dictionaries (Macdonell 1893; Monier-Williams 1899) were downloaded from the *Cologne Digital Sanskrit Dictionaries* website at Cologne University.[31] Macdonell (1893) contains slightly less than 21,000 entries while Monier-Williams (1899) has almost 290,000 entries.[32] The proportion of verbs and nouns is roughly the same in both dictionaries, although the nouns increase their share of the total vocabulary slightly (by 8%) going from the smaller to the larger dictionary, while the share of verbs decreases more noticeably (by almost 19%), see Table 3. Similarly, when the Intercontinental Dictionary Series list (Borin, Comrie, et al. 2013) was extended from the original 1,310 entries to 1,460 entries for use in the Loanword Typology project (Haspelmath & Tadmor 2009), the VNR dropped from 0.41 to 0.37, as 112 of the added 150 entries were nouns and only 12 verbs. Again, obviously further research is needed here.

Table 3. Verbs and nouns in two Sanskrit dictionaries

	Entries	Verbs	Nouns	VNR	ΔV	ΔN
Macdonell 1893	20,749	~935	~9,600	0.098	0	0
Monier-Williams 1899	287,443	~10,500	~143,300	0.074	−0.186	+0.080

information for replication of studies, as well as actually having them replicated by more than one researcher.

31. https://www.sanskrit-lexicon.uni-koeln.de/ (accessed on 2020-12-31)

32. Both counts calculated by automatic processing of the markup as documented in the metadata files included in the downloadable versions of the dictionaries.

3.3.5.2 *Verbs, nouns and other parts of speech in corpora and dictionaries*

The verb-to-noun ratio as calculated both from a conventional dictionary and from a corpus will vary considerably, e.g. depending on if regularly derived deverbal nouns are listed in the dictionary – or lemmatized as the verb (the basis of the derived noun) or as nouns in the corpus – or not. For English, PWN regularly lists deverbal nouns in *-ing* as separate entries, while these are rarely provided by https://www.dictionary.com/. Corpus counts of verbs will vary according to how participles are lemmatized: as verbs or as independent lexical entries. More generally, there is also a confound – recognized as such by Polinsky (2012: 351) – introduced by the existence of different criteria for defining nouns and verbs, not only between different languages, but sometimes even in different descriptions of the same language (e.g. Munro 2005).

In the same way the lexical adjective–adverb ratio for English will vary significantly, should we wish to calculate it, depending on how the consulted dictionary or corpus annotation treats deadjectival adverbs in *-ly*.[33]

The core resource of SweFN++, Saldo, is undergoing a major POS revision for its next major version. Among other things, participles used to be considered inflectional forms of verbs – this is the analysis found in older Swedish school grammars – but the *Swedish Academy Grammar* (Teleman et al. 1999) instead recognizes participles as a separate part of speech (new to Swedish). For various theoretical and practical reasons, this is not the model adopted in Saldo and SweFN++, where participles are now considered to be deverbal adjectives. Of course, this does not change the VNR as calculated based on Saldo[34] – the number of verb lemmas and the number of noun lemmas remain the same – but it will lower it in corpora annotated using the new POS system compared to those where the present system is used, when participles are no longer lemmatized as verbs.

3.3.5.3 *Headedness, verb-to-noun ratio, and support verb constructions*

Polinsky & Magyar (2020) find that HF languages tend to have lower VNR than both SVO and HI languages. They then tie this tendency to strategies for deriving new verbs in a language. In short, they conclude that relative paucity of simplex verbs is a consequence of a language having turned head-final, not the other way around, The central idea proposed by them for the purposes of our discussion is

---

33. Again, as a rule PWN supplies these as separate entries, while https://www.dictionary.com/ does not.

34. The VNR of Swedish as calculated from Saldo is 0.09, which puts it among the verb-poor languages, even if not at all HF, tied with Hindi in Polinsky & Magyar (2020), and just below German and Tsez (both 0.14), with only Japanese and Persian below it at 0.04.

the suggestion that languages poor in simplex verbs will turn to employing SVCs for expressing verbal concepts.

We note in passing that it would be good to have more information about when a language is to be considered verb-poor. How many verbal concepts should we expect a normal (or average) language to express, or how many verbal concepts should we minimally expect a language to have? In our context, a very natural place to reach for the beginnings of an answer is of course the FrameNet, and say that the number of FN frames (about 1,200; Gilardi & Baker 2018) should constitute a communicative minimum for any language.

Thus, if we find much fewer verb lexemes than 1,200 in a language – like the 100 or so in Kalam or Farsi – we should ask ourselves where all the verbs have gone. Then again, the number of verb lexemes does not equal the number of *verb senses*, which is what is at issue here; one logically possible way of dealing with verb paucity in a language could be by resorting to extensive colexification. Even for English, which is not verb-poor, lexicographic practice recognizes 60–70 senses of the verb *break* (see Section 3.2).

Polinsky & Magyar (2020) do not consider the possibility that verb-poor languages increase verbal colexification. Instead they conjecture that such languages will employ more SVCs and connect this to the structuring of information content in linguistic expressions, and specifically a response to a desire on the part of the speaker to distribute information content more evenly over clauses and not to force the listener to wait for the clause-final verb in order to understand which kind of situation is depicted. This problem is solved by resorting to SVCs, which ensure that the "real" predicate will arrive slightly earlier than with a single clause-final lexical verb.

As opposed to the relationship between headedness and VNR, where they present calculations to back up their claims, it should be noted that the correlation between low VNR and SVCs is only anecdotically supported.

Savary et al. (2018) provide a rich array of statistics for a corpus annotated for verbal MWEs in 18 languages. For several of the languages they provide statistics for SVCs. The tendency is not completely clear, but we find some confirmation of the findings of Polinsky & Magyar (2020), in that the head-final languages in the corpus for which they provide SVC statistics – Hungarian and Turkish – show high proportions of SVCs in relation to all verbal MWEs: 56% for Hungarian and 42% for Turkish. On the other hand we saw above that Hungarian does not seem to be very verb-poor (or at least this should be investigated more thoroughly).

However, there are quite a few surprises in these data: the Semitic languages Maltese and Hebrew have very different SVC proportions: 54% and 21% (of all verbal MWEs), respectively. Several Slavic languages are represented in the corpus, and again the numbers are not what we would expect if genealogy is the main

determinant for the occurrence of SVCs. The South Slavic – hence very closely related – languages Slovene and Bulgarian show 10% and 21%, respectively, and the West Slavic languages Polish and Czech 40% and 20%, respectively. Another language family represented by several languages is Romance, with French (33%), Italian (43%), Portuguese (62%), Romanian (25%), and Spanish (26%). The outliers in this dataset are Greek and Portuguese (64% and 62%), and at the other extreme German and Swedish (7% and 9%). All these languages are SVO, and the differences among them do not correlate with different positions on the HF–HI scale, as far as I can see. But, as Polinsky & Magyar (2020) are careful to point out, SVCs can and will appear in languages of all types; there is no two-way implication involved.

How reliable are these numbers? There are indications in the literature that linguistic analysis, including lexical analysis, is a skill that requires both long training and extensive practical experience. Definitions of lexical phenomena are often not very detailed; instead the analysis has strong elements of expert knowledge, the kind of tacit knowledge acquired through long experience. Especially corpus annotation procedures vary widely, and are not always well documented.

3.3.5.4 *The diachrony of support verb constructions in verb-final languages*

In brief, the diachronic scenario depicted by Polinsky & Magyar (2020) is one where headedness – specifically the position of the verb in the clause – drives a development towards verb paucity or richness. Thus, they suggest that the development from relatively verb-poor Latin to relatively verb-rich Romance (represented in their sample by French and Spanish) is a consequence of a shift in main word order. They also note that they do not at present have the data to confirm whether a language turning more head-final is followed by shrinkage of the set of simplex verbs and a concomitant increase in SVC usage.

One language where a loss of simplex verbs is known to have occurred is Persian, which over its recorded history has seen a continuous replacement of simplex verbs with support verbs, with the result that the number of simplex (SWE) verbs in the modern language (Farsi) is slightly above 100 (Mohammad & Karimi 1992).

However, as confirmation of Polinsky & Magyar's (2020) hypothesis it is problematic, since Persian has been OV throughout its recorded history (and before, since Proto-Indo-European is also normally reconstructed as OV), and has if anything become less head-final with time. At the same time the number of lexical verbs has shrunk and the number of SVCs has increased consistently, starting already in Proto-Indo-Iranian (since both Sanskrit and Old Persian already exhibit this construction; Ciancaglini 2011). In fact, modern Indo-Aryan languages tend to be more head-final than modern Persian, which despite this shows a greater variety in its SVCs than the former.

Notably, the VNR has not changed from Sanskrit (Old Indo-Aryan) to Hindi (New Indo-Aryan; VNR 0.09); see Table 3. Classical Sanskrit (the language of the two dictionaries) already had a large number of SVCs (Ittzés 2015), and it is not known if their number has increased in Modern Indo-Aryan languages. Like in the case of Persian, most varieties of Indo-Aryan have remained steadfastly OV throughout history.

The increase in SVCs in Persian seems to have taken off coincident with a major simplification in both nominal and verbal morphology which marks the transition from the Old to the Middle Persian stage (Maggi & Orsatti 2018), and which notably removes much dependent-marking on verb arguments, thus shifting clausal information content towards the end of the clause, since the role of an argument is no longer signalled by its morphology, but must be understood from the semantics of the verb, and hence is available only at the end of the clause. On the other hand, the transition from Old to Middle Indo-Aryan is also one of massive loss of inflectional morphology, but one which possibly has not then been accompanied by an increase in SVCs.

In this connection I note that Hungarian (HF) has a *very* rich nominal inflectional morphology allowing for specification in almost painful detail of the roles of nominal dependents of various parts of clauses and phrases, and notably of verbs. This fits well with Hungarian not being so verb-poor, after all.

In the wider context of typological aspects of MWEs, the correlation that Polinsky & Magyar (2020) point to could arguably be seen as a special case of a more general "design feature" difference among languages, saying nothing now about its possible origins. A concrete example: Icelandic and Finnish are sociolinguistically similar, in that both languages reach for a rich native word-formation machinery rather than resort to borrowing for introducing new lexemes in their vocabulary. Even though both languages possess both compounding and derivation as word-formation devices, there is a systematic difference in their usage, which arguably reflects typological differences between them. A Finnish derivation will typically correspond to an Icelandic compound,[35] e.g. *tark-asta-ja* [exact/precise-VBZ-AGENT] (Finnish) : *eftir-lit-s-maður* [after-sight-SG.GEN-person] (Icelandic), both meaning 'inspector, overseer'.[36]

35. Icelandic also has many SVCs.

36. Less puristic Swedish has a loanword: *inspektör* (or *kontrollant*).

3.4 What kinds of MWEs are there?

Several of the most cited works on MWEs in LT and phraseology in linguistics provide classifications of MWEs. As already noted above, these studies are often contrastive rather than typological in scope, and furthermore based on a small number of predominantly European languages.

Some of the studies presented in the international PARSEME initiative represent steps in the right direction, although their database is still far from representative of the world's linguistic diversity. The PARSEME corpus of verbal MWEs described by Savary et al. (2018) contains annotated data for 18 languages, out of which 14 are Indo-European, 2 Afroasiatic, 1 Finno-Ugric, and 1 Turkic. In addition, there is a strong geographical skew in the dataset. All except one language – Farsi – are either European or Mediterranean or both. From a strictly typological point of view there is arguably a risk that this corpus represents at most 1.5 datapoints in a global perspective, out of a few hundred that would be required for a first unbiased survey of the space of variation in this domain (Bakker 2011). Both Europe and the Mediterranean have been suggested as linguistic areas (Haspelmath 2011a; Sansó 2011). This would exclude only Farsi from possible contact influences,[37] but on the other hand Farsi is Indo-European like most of the other languages in the corpus. How typologically diverse this corpus is likely to be is dependent i.a. on how MWEs emerge in languages: do they find their way into the language by inheritance (genealogically), by borrowing/calquing (by language contact),[38] or spontaneously, by language-internal mechanisms? There is a chicken-and-egg problem here, in that we need relevant data from a broad range of languages in order to even begin to establish this. Ideally, in order to throw light on the mechanisms by which languages acquire MWEs, we would also require diachronic studies.

Thus it would be a desirable goal for the future to extend the database more systematically, considering typological sampling requirements, and also with a view to survey the actual range of MWE types in the languages of the world. We should not exclude a priori that the MWE types posited on the basis of examination of a small number of primarily European languages may be unusual in a global perspective.[39]

---

37. Although Farsi has been under influence from Arabic – which is included in the corpus in the form of Maltese – for a very long time.

38. At least some kinds of MWEs seem to have spread over Europe through language contact (see Piirainen 2012), and Matras (2007: 47ff) mentions that support verb constructions are a cross-linguistically common strategy for accommodating borrowed verbs in languages, although to be clear: he does not say that the construction itself tends to be borrowed.

39. As seen in Example 5, the "Dances with wolves" noun formation model – when a clause is used as a noun, but retains its clause-internal grammar – known from popular accounts of North American languages, is also prevalent in Kalam (New Guinea).

The history of linguistic typology is a history of constant surprise – of constant experiences of a "sense of wonder", to borrow a term from science fiction criticism – large and small. It is now known that several linguistic features which are implicitly taken for granted as self-evidently belonging to the most basic stock items of linguistic description are actually by and large confined to Europe or much overrepresented among European languages (of more than one family, to boot), e.g., indefinite and definite articles, relative clauses using a relative pronoun, comparison constructed using a particle and a special comparative adjective form, comitative-instrumental syncretism, participial passive, and several others (see Haspelmath 2011a).

Some conceptual issues need to be clarified first. Is it meaningful to embark on an investigation of MWEs in a broad cross-linguistic perpective, or are MWEs simply an epiphenomenon caused by idiosyncratic quirks of a few orthographic systems?[40] This is not to deny that the practical technical problems of segmentation of written texts in LT are not interesting or difficult (Chiarcos et al. 2012; Dridan & Oepen 2012), only that they are not considered to be central to linguistic typology. Typologists enquire about the possible and its converse, the impossible (or rather imaginable but so far unattested). Given the way that languages seem to work, the main preoccupation of typologists is with (statistical) distributions and correlations of (paradigmatic) alternatives in languages: For instance, how many languages exhibit SO (subject before object) as opposed to OS basic word order, and how highly do the SO and OS orders correlate with other ordering regularities in language? See the various chapters in the *World atlas of language structures* (WALS; Dryer & Haspelmath 2013). As has been pointed out repeatedly by Haspelmath (2010), a robust conceptual apparatus allowing the comparison of language-specific phenomena is an indispensable prerequisite for all typological work. In other words, a clear definition of MWEs is needed in order to look for them in language descriptions and language data, in the same way that we need to know beforehand what a face looks like in order to find the hidden faces in the image puzzle reproduced in Figure 1.[41]

---

40. The vast majority of the world's languages are not written, so that a "linguistic" generalization which holds only for written languages is much less valuable than one that holds for all languages.

41. Image source: https://chroniclingamerica.loc.gov/lccn/sn85066387/1913-06-29/ed-1/seq-40/. This image is in the public domain: "The Library of Congress believes that the newspapers in Chronicling America are in the public domain or have no known copyright restrictions. Newspapers published in the United States more than 95 years ago are in the public domain in their entirety." (https://chroniclingamerica.loc.gov/about/)

Figure 1. A classical hidden-objects picture puzzle from *The San Francisco call*, 29 June, 1913

3.5 Where do we find cross-linguistic MWE data?

A central issue concerns the availability of empirical data for cross-linguistic comparison of MWEs as a linguistic phenomenon. As already mentioned, linguistic typology works with large language samples, preferably 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 (the typical scenario) to standard reference grammars (a rare treat). Tailor-made questionnaires focusing on specific features are also common in these investigations. Such secondary sources seldom contain information on MWE phenomena (see, e.g., Schultze-Berndt 2006: 371ff). One of the most widely used typological databases, WALS, covers close to 200 linguistic features and almost 2,700 languages,[42] but there are no obvious features relevant to MWEs.

42. Although the database as a whole is quite sparse: even though WALS reports values for a total of 192 linguistic features in 2,662 languages, in reality most cells in the resulting matrix are empty. In version 2020 of the dataset available for download from http://wals.info/download, out of a total of 511,104 cells, no less than 434,625 are empty, meaning that less than 15% of the potential values are actually recorded in the database.

Another conceivable source for information on MWEs would be dictionaries, but in practice, such information is absent or very hard to find even in large monolingual reference dictionaries.[43] Even when they are treated in traditional dictionaries, MWEs may be provided only inside "proper" (i.e. SWE) lexical entries – often with unclear or unstated principles determining under which component word they should be listed – and not given as headwords in their own right.[44]

As already mentioned, the greatest conundrum is perhaps the conceptual foundation: exactly what are we looking for?

Some specific constructions which conform to Baldwin & Kim's (2010) definition cited in Section 3.1 above – but possibly not to all construals of MWEs in the LT literature – have generated a considerable number of publications in linguistic typology, amassing data from many languages. This concerns constructions such as *compounding* (Lieber & Štekauer 2009; Bauer 2009) and *incorporation* (Mithun 1984; Amith 2002), as well as what is often called *complex predicates* in the literature (e.g., Bowern 2008). Among these we find *serial verb constructions* (Aikhenvald & Dixon 2006), and *light* or *support verb constructions* (Schultze-Berndt 2006; Butt 2010). This body of work is obviously relevant to our question. Both compounding and incorporation are characterized as mechanisms invoked when "[s]ome entity, quality, or activity is recognized sufficiently often to be considered name-worthy in its own right" (Mithun 1984: 848), or "culturally salient" (Amith 2002: 237), which would at a minimum constitute "lexical idiomaticity" and consequently be relevant to our goals. On the other hand, MWE adverbs and function items are generally not considered in the literature, which is notable given their prevalence in Saldo (as shown in Table 1).

Even though these studies generally fall far short of the sampling standards expressed by Bakker (2011) – for instance, Lieber & Štekauer (2009) present studies on compounding in 17 languages, out of which 7 are Indo-European – they are still significantly ahead of the state of the art in typical LT papers on MWEs. Thus, Sag et al. (2002) deal exclusively with English, and while occasionally nodding at other languages, it is still clear that English is the focus even of the survey of the field presented in Baldwin & Kim (2010). The already mentioned study by Savary et al. (2019) is a step in the right direction. They investigate verbal MWEs

---

43. The importance of MWEs is arguably underrated in traditional lexicography, and this probably also spills over into linguistics in the form of a bias. Thus, the "meanings" studied by Majid et al. (2015) are assumed to be expressed preferably by single words, as becomes clear from their description of the coding procedure applied to their experimental data (Majid et al. 2015: 7).

44. Incidentally, this may be one reason for the low presence of MWEs in Saldo, mentioned earlier, since an important source of Saldo entries was a list of headwords from a large conventional Swedish dictionary, but not any additional information from the entries of that dictionary.

in five languages, even if their characterization of the investigated languages as "typologically different" should be understood against the background of the state of the art in LT; all are European languages, and four of them are Indo-European (German, Greek, Polish and Portuguese), although representing four separate primary branches of this language family, and the fifth is the European language isolate Basque.

4. Taking stock: Towards a typology of MWEs?

The present chapter describes and motivates the approach taken in SweFN++ to the description of multiword expressions, and the theoretical and methodological questions raised in connection with this work, especially if we would like to think of it in the wider context of lexical typology.

Against the background presented above: 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 LT as opposed to (typological) linguistics?

For reasons given above in Section 3.1, orthographic words cannot be used in a language-independent characterization of MWEs. We should rather be striving for something similar to Haspelmath's (2015: 296) 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', i.e. "a construct that may occur on its own as a complete utterance" – since "all lexical items have citation forms that are free constructs" (Haspelmath 2011b: 70).

Thus, we are looking for constructs involving more than one lexeme, where the behavior of the whole deviates from compositionality (which also needs to be defined in a language-independent way, of course; cf. Section 3.2 above). From a typological point of view, we wish to cast as wide a net as possible, i.e. without too many preconceptions as to what kinds of constructs we will uncover.

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., LT or lexicography. Presumably, all language-specific MWEs will also be typological MLEs (but the opposite will not hold; e.g., German compounds are definitely MLEs but may or may not be considered MWEs).

There will be difficult cases. If "lexeme" is equated with 'conventional lexical entry', English words such as *ongoing* or *undertake* will have to be considered MLEs, since they are non-compositional and consist of more than one lexeme. On the other hand, the long "sentence-words" of Eskimo-Aleut or Wakashan languages, containing only one independent lexeme and a long series of suffixes, even when lexicalized will not count as MLEs according to this definition. Rather, they will present themselves as extreme versions of derivational morphology Examples (6) and (7) – as seen in the Finnish *-sto/-stö* formations mentioned earlier.

(6) Inuktitut (ike) (Eskimo-Aleut, Canada; Dorais 2017: 148)
aupalut-si-guti
something.red-TO.MAKE.IT.SO-THAT.IS.USED.TO
'That is used to redden something (=lipstick)'

(7) Nuu-chah-nulth (nuk) (Wakashan, Canada; Mithun 1984: 888)
a. č'apac-o'ał
canoe-PERCEIVE
'see a canoe'
b. č'apac-nak
canoe-POSSESSING
'having a canoe'

As mentioned in Section 3.5, even secondary data on MWEs are largely lacking. Still, a reasonable first step would be to survey the literature on likely kinds of MWEs (compounds, incorporation, complex predicates, full-clause nominalizations, etc.), in order to formulate some tentative generalizations and research questions.

In a longer perspective, the methods developed in LT for identifying MWE candidates in corpora (see, e.g., Pecina 2010) or even methods for unsupervised word segmentation (e.g. Hewlett & Cohen 2011), or more generally for terminology mining (e.g. Wermter & Hahn 2005), could potentially be of great help in taking this research further. In particular, we would expect such approaches to provide tools allowing us to treat conventionalization and lexicalization as gradient rather than categorical phenomena.

Even if grammars or lexicons do not mention MLEs, this information may be hidden in the text collections often accompanying descriptive grammars. These text materials tend to be modest in size, so that we would need good methods for uncovering idiomaticity on the basis of small corpora. Incidentally, we note that developing a linguistic 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.

At the same time, the deeper understanding of MWEs resulting from such research should hopefully serve to increase our ability to cope with language-specific challenges and opportunities arising from dealing with MWEs in practical LT applications.

Funding

The research presented here has been supported by the Swedish Research Council (the projects *Digital areal linguistics*: grant 2009–01448; *Swedish FrameNet++*: grant 2010–6013; *South Asia as a linguistic area?*: grant 2014–00969; *The National Swedish Language Bank*: grant 2017–00626), and by the University of Gothenburg through its support of the Centre for Language Technology and of Språkbanken Text.

References

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

Aikhenvald, Alexandra. 2007. Typological distinctions in word-formation. In Timothy Shopen (ed.), *Language typology and syntactic description. Volume III: Grammatical categories and the lexicon*, 2nd edn., 1–65. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511618437.001

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

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

Amith, Jonathan D. 2002. What's in a word? The *whys* and *what fors* of a Nahuatl dictionary. In William Frawley, Kenneth C. Hill & Pamela Munro (eds.), *Making dictionaries: Preserving indigenous languages of the Americas*, 219–258. Berkeley: University of California Press.

Baayen, R. Harald. 2001. *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers. https://doi.org/10.1007/978-94-010-0844-0

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

Baldwin, Timothy, Colin Bannard, Takaaki Tanaka & Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of MWE 2003*, 89–96. Sapporo: ACL. https://doi.org/10.3115/1119282.1119294

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

Bauer, Laurie. 2009. Typology of compounds. In Rochelle Lieber & Pavol Štekauer (eds.), *The Oxford handbook of compounding*, 343–356. Oxford: Oxford University Press.

Bickel, Balthasar & Fernando Zúñiga. 2017. The 'word' in polysynthetic languages: Phonological and syntactic challenges. In Michael Fortescue, Marianne Mithun & Nicholas Evans (eds.), *The Oxford handbook of polysynthesis*, 158–185. Oxford: Oxford University Press.

Borin, Lars, Bernard Comrie & Anju Saxena. 2013. The Intercontinental Dictionary Series – a rich and principled database for language comparison. In Lars Borin & Anju Saxena (eds.), *Approaches to measuring linguistic differences*, 285–302. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110305258.285

Borin, Lars, Markus Forsberg & Lennart Lönngren. 2013. SALDO: A touch of yin to WordNet's yang. *Language Resources and Evaluation* 47(4): 1191–1211. https://doi.org/10.1007/s10579-013-9233-4

Bowern, Claire. 2008. The diachrony of complex predicates. *Diachronica* 25(2): 161–185. https://doi.org/10.1075/dia.25.2.03bow

Burger, Harald, Dmitrij Dobrovol'skij, Peter Kühn & Neal R. Norrick (eds.). 2007. *Phraseology: An international handbook of contemporary research (2 volumes)*. Berlin: Walter de Gruyter. https://doi.org/10.1515/9783110190762

Butt, Miriam. 2010. The light verb jungle: Still hacking away. In Mengistu Amberber, Brett Baker & Mark Harvey (eds.), *Complex predicates: Cross-linguistic perspectives on event structure*, 48–78. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511712234.004

Chiarcos, Christian, Julia Ritz & Manfred Stede. 2012. *By all these lovely tokens…* Merging conflicting tokenizations. *Language Resources and Evaluation* 46(1): 53–74. https://doi.org/10.1007/s10579-011-9161-0

Ciancaglini, Claudia A. 2011. The formation of the periphrastic verbs in Persian and neighbouring languages. In Mauro Maggi & Paola Orsatti (eds.), *The Persian language in history*, 3–31. Wiesbaden: Reichert.

Constant, Mathieu, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner & Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics* 43(4): 837–892. https://doi.org/10.1162/COLI_a_00302

Czuczor, Gergely & János Fogarasi. 1862. *A magyar nyelv szótára* [Hungarian dictionary]. Pest: Emich Gusztáv Magyar akadémiai nyomdász.

Dixon, R. M. W. 2006. Serial verb constructions: Conspectus and coda. In Alexandra Aikhenvald & R. M. W. Dixon (eds.), *Serial verb constructions: A cross-linguistic typology*, 338–350. Oxford: Oxford University Press.

Dorais, Louis-Jacques. 2017. The lexicon in polysynthetic languages. In Michael Fortescue, Marianne Mithun & Nicholas Evans (eds.), *The Oxford handbook of polysynthesis*, 135–157. Oxford: Oxford University Press.

Dridan, Rebecca & Stephan Oepen. 2012. Tokenization: Returning to a long solved problem – a survey, contrastive experiment, recommendations, and toolkit. In *Proceedings of ACL 2012*, 378–382. Jeju: ACL.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The world atlas of language structures online*. Jena: Max Planck Institute for the Science of Human History.

Eberhard, David M., Gary F. Simons & Charles D. Fennig (eds.). 2021. *Ethnologue: Languages of the world*. 24th edn. Dallas: SIL International.

Erk, Katrin. 2010. What is word meaning, really? (And how can distributional models help us describe it?) In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, 17–26. Uppsala: ACL.

Evans, Nicholas. 2011. Semantic typology. In Jae Jung Song (ed.), *The Oxford handbook of linguistic typology*, 504–533. Oxford: Oxford University Press.

Fortescue, Michael, Marianne Mithun & Nicholas Evans (eds.). 2017. *The Oxford handbook of polysynthesis*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199683208.001.0001

François, Alexandre. 2008. Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In Martine Vanhove (ed.), *From polysemy to semantic change: Towards a typology of lexical semantic associations*, 163–215. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.106.09fra

Gantar, Polina, Carla Parra Escartín & Héctor Martinez Alonso. 2019. Multiword expressions: Between lexicography and NLP. *International Journal of Lexicography* 32(2): 138–162. https://doi.org/10.1093/ijl/ecy012

Gibbs, Raymond W., Jr, Nandini P. Nayak & Cooper Cutting. 1989. How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of Memory and Language* 28: 576–593. https://doi.org/10.1016/0749-596X(89)90014-4

Gilardi, Luca & Collin Baker. 2018. Learning to align across languages: Toward Multilingual Frame Net. In *Proceedings of the International FrameNet workshop at LREC 2018: Multilingual framenets and constructicons*, 13–22. Miyazaki: ELRA.

Hakulinen, Auli, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen & Irja Alho. 2004. *Iso suomen kielioppi* [The big Finnish grammar]. Online version at http://scripta.kotus.fi/visk/, accessed on 2021-04-22. Helsinki: Suomalaisen Kirjallisuuden Seura.

Hammarström, Harald, Robert Forkel & Martin Haspelmath (eds.). 2020. *Glottolog 4.3*. Jena: Max Planck Institute for the Science of Human History. https://doi.org/10.5281/zenodo.4061162

Hanks, Patrick. 2000. Do word meanings exist? *Computers and the Humanities* 34(1–2): 205–215. https://doi.org/10.1023/A:1002471322828

Hanks, Patrick. 2002. Mapping meaning onto use. In Marie-Hélène Corréard (ed.), *Lexicography and natural language processing: A Festschrift in honour of B.T.S. Atkins*, 156–198. Grenoble: EURALEX.

Hanks, Patrick. 2013. *Lexical analysis: Norms and exploitations*. Cambridge: MIT Press. https://doi.org/10.7551/mitpress/9780262018579.001.0001

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3): 663–687. https://doi.org/10.1353/lan.2010.0021

Haspelmath, Martin. 2011a. The European linguistic area: Standard Average European. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language typology and language universals: An international handbook*. Vol. 2, 1492–1510. Berlin: Walter De Gruyter.

Haspelmath, Martin. 2011b. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(1): 31–80. https://doi.org/10.1515/flin.2011.002

Haspelmath, Martin. 2015. The serial verb construction: Comparative concept and cross-linguistic generalizations. *Language and Linguistics* 17(3): 291–319. https://doi.org/10.1177/2397002215626895

Haspelmath, Martin & Uri Tadmor (eds.). 2009. *Loanwords in the world's languages: A comparative handbook*. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110218442

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

Hoffmann, Thomas & Graeme Trousdale (eds.). 2013. *The Oxford handbook of construction grammar*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195396683.001.0001

Ittzés, Máté. 2015. Light verb constructions in Vedic. *Manas* 2(1).

Jackendoff, Ray. 1997. *The architecture of the language faculty*. Cambridge: MIT Press.
Jensen, John T. 1990. *Morphology: Word structure in generative grammar*. Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.70
Kilgarriff, Adam. 1997. I don't believe in word senses. *Computers and the Humanities* 31(2): 91–113. https://doi.org/10.1023/A:1000583911091
Koptjevskaja-Tamm, Maria, Martine Vanhove & Peter Koch. 2007. Typological approaches to lexical semantics. *Linguistic Typology* 11: 159–185. https://doi.org/10.1515/LINGTY.2007.013
Lass, Roger. 1978. Mapping constraints in phonological reconstruction: On climbing down trees without falling out of them. In Jacek Fisiak (ed.), *Recent developments in historical phonology*, 245–286. Berlin: De Gruyter. https://doi.org/10.1515/9783110810929.245
Lieber, Rochelle & Pavol Štekauer (eds.). 2009. *The Oxford handbook of compounding*. Oxford: Oxford University Press.
Lyngfelt, Benjamin, Linnéa Bäckström, Lars Borin, Anna Ehrlemark & Rudolf Rydstedt. 2018. Constructicography at work: Theory meets practice in the Swedish constructicon. In Benjamin Lyngfelt, Lars Borin, Kyoko Ohara & Tiago Timponi Torrent (eds.), *Constructicography: Constructicon development across languages*, 41–106. Amsterdam: John Benjamins. https://doi.org/10.1075/cal.22.03lyn
Macdonell, Arthur A. 1893. *A Sanskrit-English dictionary: Being a practical handbook with transliteration, accentuation, and etymological analysis throughout*. London: Longmans, Green.
Maggi, Mauro & Paola Orsatti. 2018. From Old to New Persian. In Mauro Maggi & Paola Orsatti (eds.), *The Oxford handbook of Persian linguistics*, 7–51. Oxford: Oxford University Press.
Majid, Asifa, Fiona Jordan & Michael Dunn. 2015. Semantic systems in closely related languages. *Language Sciences* 49: 1–18. https://doi.org/10.1016/j.langsci.2014.11.002
Markantonatou, Stella, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.). 2018. *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*. Berlin: Language Science Press.
Matras, Yaron. 2007. The borrowability of structural categories. In Yaron Matras & Jeanette Sakel (eds.), *Grammatical borrowing in cross-linguistic perspective*, 31–73. Berlin: Walter de Gruyter. https://doi.org/10.1515/9783110199192.31
Mithun, Marianne. 1984. The evolution of noun incorporation. *Language* 60(4): 847–894. https://doi.org/10.1353/lan.1984.0038
Mithun, Marianne. 2009. Polysynthesis in the Arctic. In Marc-Antoine Mahieu & Nicole Tersis (eds.), *Variations on polysynthesis: The Eskimo-Aleut languages*, 3–18. Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.86.01pol
Mohammad, Jan & Simin Karimi. 1992. 'Light' verbs are taking over: Complex verbs in Persian. In *Proceedings of WECOL 1992*, 195–212. Fresno: Dept. of Linguistics, California State University, Fresno.
Monier-Williams, Monier. 1899. *A Sanskrit-English dictionary: Etymologically and philologically arranged with special reference to cognate Indo-European languages*. Oxford: The Clarendon Press.
Munro, Pamela. 2005. From parts of speech to the grammar. *Studies in Language* 30(2): 307–349. https://doi.org/10.1075/sl.30.2.07mun
Nasr, Alexis, Carlos Ramisch, José Deulofeu & André Valli. 2015. Joint dependency parsing and multiword expression tokenization. In *Proceedings of ACL/ IJCNLP 2015*, 1116–1126. Beijing: ACL. https://doi.org/10.3115/v1/P15-1108
Nida, Eugene A. 1949. *Morphology: The descriptive analysis of words*. Ann Arbor: University of Michigan Press.

Nivre, Joakim & Jens Nilsson. 2004. Multiword units in syntactic parsing. In *Proceedings of MEMURA at LREC 2004*, 39–46. Lisbon: ELRA.

Nunberg, Geoffrey, Ivan A. Sag & Thomas Wasow. 1994. Idioms. *Language* 70(3): 491–538. https://doi.org/10.1353/lan.1994.0007

Parmentier, Yannick & Jakub Waszczuk (eds.). 2019. *Representation and parsing of multiword expressions: Current trends*. Berlin: Language Science Press.

Pawley, Andrew. 1993. A language which defies description by ordinary means. In William A. Foley (ed.), *The role of theory in language description*, 87–129. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110872835.87

Pecina, Pavel. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation* 44(1–2): 137–158. https://doi.org/10.1007/s10579-009-9101-4

Piirainen, Elisabeth. 2012. *Widespread idioms in Europe and beyond: Toward a lexicon of common figurative units*. New York: Peter Lang. https://doi.org/10.3726/978-1-4539-0845-7

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

Polinsky, Maria & Lilla Magyar. 2020. Headedness and the lexicon: The case of verb-to-noun ratios. *Languages* 5(1/9): 1–25. https://doi.org/10.3390/languages5010009

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Alexander Gelbukh (ed.), *Computational linguistics and intelligent text processing: Third international conference: Cicling-2002*, 1–15. Berlin: Springer. https://doi.org/10.1007/3-540-45715-1_1

Sailer, Manfred & Stella Markantonatou (eds.). 2018. *Multiword expressions: Insights from a multi-lingual perspective*. Berlin: Language Science Press.

Sansó, Andrea. 2011. Mediterranean languages. In Bernd Kortmann & Johan van der Auwera (eds.), *The languages and linguistics of Europe: A comprehensive guide*, 341–356. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110220261.341

Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87–147. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.1471591

Savary, Agata, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta & Voula Giouli. 2019. Literal occurrences of multiword expressions: Rare birds that cause a stir. *The Prague Bulletin of Mathematical Linguistics* (112): 5–54. https://doi.org/10.2478/pralin-2019-0001

Schulte im Walde, Sabine & Eva Smolka (eds.). 2020. *The role of constituents in multiword expressions: An interdisciplinary, cross-lingual perspective*. Berlin: Language Science Press.

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

Silveira, Natalia & Christopher D. Manning. 2015. Does Universal Dependencies need a parsing representation? An investigation of English. In *Proceedings of Depling 2015*, 310–319. Uppsala: ACL.

Skorik, Pëtr J. 1961. *Grammatika čukotskogo jazyka: Čast' pervaja* [Chukchi grammar: Part I]. Moscow: Izdatel'stvo Akademii Nauk SSSR.

Taljard, Elsabé & Sonja E. Bosch. 2006. A comparison of approaches to word class tagging: Disjunctively vs. conjunctively written Bantu languages. *Nordic Journal of African Studies* 15(4): 428–442.

Tanaka, Takaaki & Timothy Baldwin. 2003. Noun-noun compound machine translation: A feasibility study on shallow processing. In *Proceedings of MWE 2003*, 17–24. Sapporo: ACL. https://doi.org/10.3115/1119282.1119285

Teleman, Ulf, Staffan Hellberg & Erik Andersson. 1999. *Svenska Akademiens grammatik* [The Swedish Academy grammar]. Stockholm: Norstedts.

van der Auwera, Johan. 2012. From contrastive linguistics to linguistic typology. *Languages in Contrast* 12(1): 69–86. https://doi.org/10.1075/lic.12.1.05auw

Wermter, Joachim & Udo Hahn. 2005. Finding new terminology in very large corpora. In *Proceedings of K-CAP'05*, 137–144. Banff: ACM. https://doi.org/10.1145/1088622.1088648

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