International Journal of Lexicography, 2018, 1–25
doi: 10.1093/ijl/ecy012
Article

Article

MULTIWORD EXPRESSIONS: BETWEEN LEXICOGRAPHY AND NLP

Polona Gantar
Faculty of Arts, University of Ljubljana (apolonija.gantar@ff.uni-lj.si)

Lut Colman
Instituut voor de Nederlandse taal (lut.colman@ivdnt.org)

Carla Parra Escartín
ADAPT Centre, Dublin City University (carla.parra@adaptcentre.ie)

Héctor Martínez Alonso
Thomson Reuters Labs (hector.martinez.a@gmail.com)

Abstract

The paper aims to establish a synergy between the lexicographic and natural language processing (NLP) communities in relation to concepts and classifications of multiword expressions (MWEs), their representation in dictionaries, dictionary databases, and NLP-oriented MWE lexicons. It begins with an overview of basic MWE-related linguistic concepts and how they are reflected in the lexicographic treatment of MWEs, as well as their role in language technology. A comparison of different lexicographic and NLP classifications of MWEs is presented, with an elaboration of why different typologies are (or are not) useful for different users from both communities. The methodology for the description of MWEs in a set of dictionary databases is discussed, and the results of an analysis of the representation of MWEs based on a small sample of dictionary projects are presented. Finally, some suggestions are provided on how to improve dictionary databases in relation to MWE description and how to improve the results of NLP tasks by using existing descriptions of MWEs in dictionaries.

1. Introduction

MWEs represent an extensive part[1] of the mental lexicon of native speakers in languages in general and, as a consequence, they also appear frequently in texts that need to be processed by computers. As such, they are an important phenomenon for research in linguistics, including its practical applications, such as compilation of dictionaries, and in natural language processing (NLP), for example in the form of machine-readable MWE lexicons used

as part of NLP tools. A number of properties of MWEs, recognised both in linguistic theory and in NLP, distinguish them from single-word lexical units. Their most obvious characteristic – that they consist of two or more parts, or words[2] – has various consequences for their treatment in dictionaries and in NLP lexicons. Individual elements within MWEs can be limited to particular forms in their (morphological) paradigms. Their semantic or syntactic relations may allow only limited lexical choice, and they show semantic non-compositionality, also referred to as idiomaticity.

Both lexicographic and NLP research communities have demonstrated considerable interest in MWEs. There are some good practices in lexicographic treatment of fixed expressions and their presentation of idiomaticity in specialised dictionaries (cf. ODCIE, SČFI). Lexicographic works on MWEs are typically available in the print medium, however, already in the 1990s DEFI (Michiels 1999) and COMPASS (Breidt et al. 1996) were early attempts to establish bridges between lexicography and NLP. The DEFI project aimed at creating an online reading comprehension tool that would offer translations between English and French of both single words and multiword units. To this end, the project team developed a matching algorithm that could also handle MWE recognition, meaning assignment, and translation selection. Besides the NLP component, the tool relied heavily on lexical data, namely dictionaries, thesauri, and a hand-made bilingual database (Michiels and Dufour 1998, Michiels 2000). Researchers of the COMPASS project chose to use the IDAREX finite-state formalism for the formal description of English, German and French idioms (Breidt et al. 1996, Segond and Breidt 1995). This formalism allowed them to control the morphological, lexical, and structural differences in idioms, as well as the modification of their components. Despite these first attempts at harnessing lexicographic work and NLP formalisms to process MWEs, their complex nature remains an obstacle for their treatment in both fields, as the reader will see in the remainder of this paper. Further proof of this is the vast body of publications devoted to the analysis, processing, and treatment of MWEs both in lexicography and in NLP.

With lexicography, the identification of a string of words as a lexical unit, the basic unit of meaning, is usually considered the primary task in relation to MWEs. Lexicographers must also determine specific lexicographic information attributed to MWEs, their treatment within the dictionary macrostructure and — due to their multiword nature — produce or enable user-friendly search strategies, which is considered a particularly challenging part of traditional lexicographers' work. The difference between traditional paper dictionaries, dictionary databases, and MWE lexicons should be emphasised here, as well as the fact that even in machine readable dictionary databases, information was not organised primarily with NLP tasks in mind, but rather to support dictionary compilation.

Lexicographers use a variety of methods and standards during the compilation of lexicographic works that include MWEs. MWEs can be treated at different levels of detail, either as independent lexical units (headwords) or as subordinate units under single-word headwords, or even as part of illustrative material. In traditional dictionaries, the internal formal structure is usually not considered at all, or at least not in the form that is immediately applicable in NLP. However, although the primary goal of lexicographic treatment of MWEs is to provide dictionary users with information about their existence, meaning, and use, some dictionaries of idioms go much further than giving definitions and illustrating the use of an idiom. Notable examples are specialised dictionaries (of idioms), such as the ODCIE (Cowie et al. 1983) for English, or SČFI for Czech (Čermák et al. 2009), which do provide a systematic treatment of the internal structure of MWEs.

A crucial task in NLP is to define the formal (and semantic) features of combinations of words that enable their identification as MWEs, i.e. separate lexical units. Deviation from regular linguistic behaviour in MWEs may represent a problem for text processing, but it is helpful as the formal indicator that a particular string of words can be identified as a potential MWE. As a consequence, MWEs which language users consider to be individual lexical units based on their idiosyncratic meaning, but which basically follow regular language rules, can represent a greater challenge for NLP than lexically and syntactically idiomatic combinations. In such cases, the semantic value of an MWE is very important, and so dictionary databases containing semantic information represent a potentially valuable resource also for NLP tasks, provided that they are machine-readable.

NLP is the ideal complement for technically-enhanced lexicography, as it deals with the automatic recognition of linguistic structures and phenomena. There are many NLP tasks that can aid lexicographic work in relation to MWEs, such as their detection (arguably the most important task), estimation of compositionality (whether a phrase has a compositional reading is an active research issue in NLP, e.g. Cordeiro et al. 2016, Schulte im Walde et al. 2016, Bott and Schulte im Walde 2017, among others), finding glosses and examples for certain expressions (Kozareva and Hovy 2010), identifying synonymy and other semantic relations e.g. by using distributional models (Santus et al. 2014, Shwartz et al. 2016).

Lexicographic methods, standards, and strategies related to the description of MWEs in dictionaries, along with the problems of automatic identification of MWEs in text, were discussed at a workshop held in April 2016, organised jointly by the PARSEME (Savary et al. 2015) and ENeL COST actions.[3] The aim of the workshop was to establish a synergy between the lexicographic and NLP communities in relation to concepts and classifications of MWEs, their representation in dictionary databases and NLP-oriented MWE lexicons, and their identification and extraction from text corpora. The workshop also aimed to specify the characteristics of an ideal MWE lexicon interface from the lexicographic point of view, and to find the best ways to integrate MWE tools in realistic lexicographic workflows.

The goal of this paper is to present the findings from the workshop, particularly in terms of basic linguistic concepts, and how they are reflected in the lexicographic treatment of MWEs, as well as their role in language technology. Furthermore, in addition to describing particular databases, this paper also aims to present a model for MWE description in the dictionary database, which could be useful for developing tools for identification and automatic extraction of MWEs from text. The structure of this paper basically follows these aims. In the next section, we present linguistic concepts and methods for the description of MWEs in dictionaries, such as collocability, contiguity, idiomaticity, compositionality, figuration and fixedness. In addition, we present the treatment of MWEs in the macro- and microstructure of dictionaries, particularly in relation to the description of features determining different MWE types.

In Section 3, we use different types of MWEs as a starting point for examining how their specific properties are described in dictionaries in relation to different dictionary users and NLP researchers. We provide an overview of different classifications of MWEs used in dictionaries and analyse features that are crucial for identifying them as MWEs for both communities.

In Section 4, we discuss the methodology for describing MWEs in a number of practical applications. We present the analysis of the treatment and representation of MWEs based

on a sample of seven dictionary projects among nine received from ENeL participants prior to the joint workshop.

In Section 5, we list key NLP tasks that can aid in the lexicographic description of MWEs, and discuss lexicography as an aid for NLP applications. We conclude with a summary of possible mutually beneficial MWE presentations in future lexicographic work, and in language technology.

2. Concepts in lexicography and NLP

Atkins and Rundell (2008: 166) define MWEs as 'all the different types of phrases that have some degree of idiomatic meaning or behaviour.' This lexicographic definition is strikingly vague compared with one that is often used in NLP: 'Lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity' (Baldwin and Kim 2010: 3). While the lexicographic definition covers the features explicitly mentioned in (b), to dictionary users these features and the distinction between different types of MWEs are matters of minor concern. For lexicographers, however, it is very important to recognize MWEs by means of distinguishing features in order to identify them and decide on how and where to put them in a dictionary.

2.1 Description of MWE features

Atkins and Rundell state that there is no set of watertight criteria to validate MWEs (2008: 360), yet lexicographers use the following features, more or less intuitively, when dealing with MWEs: collocability, contiguity, idiomaticity, compositionality, figuration and fixedness.

Collocability in lexicography refers to the meaningfulness of lexical items occurring together in strings. Not all strings of words are meaningful from a lexicographic point of view. For example, a string like *drastically drop* is meaningful, whereas strings like *she put in the* and *you and* are not. Meaningful strings can be contiguous or non-contiguous. A contiguous string like *beat the record* can be interrupted by *speed*. The result is then a new contiguous expression *beat the speed record* in which *beat the . . . record* is non-contiguous.

Idiomaticity occurs when the components of an expression deviate from their regular behaviour in one way or another. Baldwin and Kim (2010) mention the lexical, syntactic, semantic, pragmatic, and/or statistical levels of idiomaticity, but we should add the morphological level to this list.[4] Lexical idiomaticity occurs when the components of an expression are not part of the conventional lexicon, as in *ad hoc* (Baldwin and Kim 2010), and when one lexical item in an expression is preferred to another by convention. For example, convention requires that one says *heavy smoker* rather than *big smoker*. An expression like *every now and then* is syntactically idiomatic, because a determiner does not precede an adverb in regular syntax. Semantic idiomaticity is probably the most familiar feature to lexicographers in validating MWEs, because it has consequences for defining policy: expressions in which the meaning cannot be obtained in the regular way, on the basis of the meanings of the components and the syntactic structure of the phrase, need to be defined in a dictionary. Idioms that are repeatedly brought up as examples of opaqueness are *kick the bucket* and *spill the beans*. An example of a pragmatically idiomatic phrase is *Good morning!* From a statistical point of view, idiomatic expressions are combinations in which the components show a strong lexical affinity, which is obvious from the high frequency of

their joint occurrence, such as *cats and dogs*. In general, levels of idiomaticity are probably the most researched and discussed topic in relation to MWEs, with many different interpretations (cf. Fillmore et al. 1988). Information about the level of idiomaticity or figurative meaning is included only implicitly in dictionaries, or is part of dictionary style guides (Atkins and Rundell 2008: 173). However, to be useful for automatic identification in NLP, a more explicit encoding of this kind of information would be required (cf. Fraser 1970).

Like idiomaticity, **compositionality** in linguistics is a feature that can occur at various levels. Ramisch defines it as 'the power of predicting the characteristics (semantics, syntax, etc.) of a group of lexemes based on the application of standard composition rules on the individual lexemes' (Ramisch 2015: 34). Lexicographers, however, use the term to refer to compositionality in meaning in particular, so it is generally used as a synonym for 'semantic idiomaticity'.

Many MWEs have a **figurative** meaning, so figuration can also be used as an indicator of an MWE. For example, an expression like *lend a hand* only makes sense in a figurative meaning.

Last but not least, **fixedness** is an important concept in the validation of MWEs. Fixed and semi-fixed phrases are important in language use and language learning, so dictionaries should record them meticulously. Fixed phrases do not allow for morphosyntactic variation or internal modification (Baldwin and Kim 2010), for example **by and larger*. Semi-fixed phrases are restricted in word order and composition, but allow some variation, such as inflection or limited lexical variation. One can say, for example, *he kicked the bucket* and one can *risk one's life* or *one's future*. Statistically relevant collocations are usually syntactically flexible.

Even with all of the above features as guidance, distinguishing non-idiomatic, compositional, or free expressions from multiword expressions is no simple task. Corpus-based linguistic and lexicographic approaches to MWEs (Moon 1998) and constructional approaches (Fillmore et al. 1988) have demonstrated that many idioms allow for far more flexibility than was initially thought, and even display a regularity and productivity of their own. Fillmore et al. demonstrate this by analysing what they call formal idioms, 'syntactic patterns dedicated to semantic and pragmatic purposes not knowable from their form alone' (1988: 505). Examples are constructions like *the X-er the Y-er*, and expressions like *let alone*. Formal idioms exhibit properties that are not fully predictable from the independently known properties of their lexical make-up and grammatical structure. However, as they can be quite productive, it is unfeasible to list them in the phrasal lexicon of the language and treat them as fixed expressions. Lexicographers have to consider ways to deal with these boundary expressions.

2.2 Theoretical frames

Lexicography has long been seen as a craft, with Atkins and Rundell (2008) even claiming that there is no such thing as 'theoretical lexicography', although lexicographers make use of concepts from, for example, lexical semantics, cognitive semantics and corpus linguistics. In contrast, Swanepoel (1994: 12) calls the 'so-called theorylessness' of practical lexicography a myth:

> Lexicographers are of necessity always meta-lexicographers: their practical descriptive activities have always been and will be informed and guided by principles or assumptions of a linguistic

> nature. These assumptions may not always be articulated, or if articulated, not strictly adhered to (which gives rise to accusations of the 'gulf' between theory and practice), and they may be eclectically constituted, i.e. they may not form a systematic or strictly coherent body of hypotheses on lexical semantic matters, but that does not diminish their status as guiding principles for lexicographic praxis. (Swanepoel 1994: 13)

Systematic theories of lexicography include Wiegand's general theory of lexicography (Wiegand 1984), and the modern theory of lexicographic functions (Bergenholtz and Tarp 2003). A comprehensive theoretical approach to the treatment of all types of MWEs in lexicography, however, is lacking, although a pronounced trend of the inclusion of collocations in dictionaries, or producing specialised collocations dictionaries can be observed in recent years (Oxford Collocations Dictionary (OCD), Macmillan Collocations Dictionary (MCD), Digitales Wörterbuch der deutschen Sprache (DWDS), also Kallas et al. 2015, etc.). Svensén (2009) mentions two different approaches in the treatment of collocations: a corpus-oriented approach, which was initiated in lexicography by John Sinclair (Sinclair 1991), and which is still prevalent in the British lexicographic tradition, and a system-oriented approach, which is characteristic of the German tradition (cf. Burger et al. 1982).

In the corpus-oriented approach (cf. Moon 1998), a collocation is a statistical phenomenon based on the frequency with which words co-occur. This approach has two disadvantages. On the one hand, if every group of words co-occurring more frequently than by chance is regarded as a collocation, then well-formed combinations that are trivial from the point of view of the language-system are included, such as *You and* or *the hotel*. On the other hand, infrequent combinations that are relevant, like proverbs, can be overlooked.

These drawbacks can be avoided with a system-oriented approach. A system-oriented approach aims to establish principles for the meaningful selection and appropriate presentation of MWEs, because the selection and presentation differs according to the type of dictionary, type of users, and the foreseen use of the dictionary. One method of describing and classifying collocations in a more system-oriented way adopts the grammatical point of view, held by, for example, Hausmann (1985) and Heid (2004). Hausmann (1985) advocates a hierarchical distinction between the base and the collocates in collocations. The base is the word a user would most likely look up to find a collocation, so the best practice to list collocations is in the entry for the base. For example:

Base + collocate:
verb + **noun** (object): *draw a **conclusion***
noun (subject) + verb: *the **day** is dawning*
adjective + **noun**: *heavy **smoker***

This may indeed be the best practice in a production-oriented collocation dictionary, which assists users in producing language, usually in the form of written text, but for reception-oriented dictionaries which include MWEs whose meaning is to be looked up, access via the collocate or both the base and the collocate may be preferable (Schubert 2014, Bahns 1996, Buendía Castro and Faber 2014).

In NLP, a similar distinction occurs between corpus-based methods (e.g. for MWE discovery) and rule-based methods using introspection (preferably supported by corpus searches, cf. Ramisch 2017, Constant et al. 2017). Similarly, there is a distinction between analysis-oriented and generation-oriented language resources and methods. The former are sometimes less precise and allow for overgeneration, because it is assumed that the

grammaticality of the analysed texts will itself exclude the majority of the ungrammatical assumptions.

Finally, another frame for describing and classifying collocations reflects a more semantic point of view. In Mel'čuk's theory (1998), for example, lexical functions are language-independent semantic constants that represent the relations among the components of a collocation. For example, a collocation like *deep sorrow* can be described as Magn (sorrow) = deep (Svensén 2009: 165). This approach was also included in the DECIDE project (Designing and Evaluating Extraction Tools for Collocations in Dictionaries and Corpora) in the 1990s (Fontenelle 1997).

2.3 Types of multiword expressions in lexicography and NLP

Atkins and Rundell (2008: 164) present a global overview of MWE types, described in dictionaries[5] as lexical items: fixed and semi-fixed phrases, phrasal idioms, proverbs and sentence-like expressions, (some) compounds and phrasal verbs. These MWE types are included in dictionaries in various ways, but in principle all of them need to fulfil the criterion that their 'meaning is more than the sum of the parts' (Atkins and Rundell 2008: 168).[6] The broad range of typologies and different decisions of lexicographers in dictionaries show that this general criterion can be interpreted in various ways. Some types of MWEs can be included in dictionaries as lexical units, even if they are in fact semantically transparent. Among these are many English compounds (e.g. *bus driver*), some verb + particle combinations, and light verb constructions. Many word combinations are very frequent and fixed in structure, but in terms of predictability of meaning, they dwell in the grey area between free combinations and MWEs (e.g. *dark chocolate)*, and their classification and inclusion in dictionaries is rather arbitrary.

The types of MWEs mentioned in Atkins and Rundell (2008: 164) are typically included in dictionaries, although not necessarily explicitly, and often under different names. We divided their initial classification into seven groups (Table 1) and compared it to other classifications proposed by Bergenholtz and Gouws (2014)[7] and Baldwin and Kim (2010).[8] The first two classifications of MWE types are devised explicitly from the lexicographical point of view, while the third is more NLP-oriented. There are other classifications which could be taken into account, but these were chosen because they take the practical view both in lexicography and NLP.

COLLOCATIONS as a type of MWE are mentioned both in relation to dictionaries, especially in the context of EFL dictionaries, and in NLP. Collocations are considered as semantically transparent (Atkins and Rundell 2008: 223), and from the lexicographic point of view they are useful mainly for achieving (native-like) proficiency in language learning, illustration of typical word use and semantic disambiguation (Sinclair 1987, 1991; Hunston and Francis 2000; Stubbs 2007; Hoey 2005), or tracing semantic changes in word use (Geeraerts 1997). From the point of view of NLP, the frequently used definition of a collocation as 'an arbitrary and recurrent word combination' (Benson 1990) emphasises the notion of 'recurrence', which is also understood as 'statistical idiomaticity' (Baldwin and Kim 2010). The notion of 'collocation' was already recognized by Firth (1957), long before the widespread ability to process large quantities of text, but it was explored only with the rise of corpus linguistics, producing a genuine revolution in lexicography (Sinclair 1991). Various statistical association measures[9] were thus designed to score the degree of 'collocability' in word combinations (Church and Hanks 1990, Manning and Schütze 1999;

Table 1. Types of MWEs in dictionaries vs. NLP classification

Atkins and Rundell (2008)	Bergenholtz and Gouws (2014)	Baldwin and Kim (2010)
I. COLLOCATIONS		
• collocations *risk one's life*	• collocations *severe criticism*	• collocations *immaculate performance*
II. FIXED PHRASES & IDIOMS		
• phrasal idioms *to have a heart of gold* • fixed phrases *ham and eggs* • similes *drunk as a lord*	• idioms *to have eyes in the back of one's head* • non-pictorial idiomatic MWE *round the clock* • twin formula *day and night* • comparative MWE *as right as rain* • MWEs from foreign languages *ad hoc* • (non)idiomatic MWEs with a unique component *to and fro* • MWEs with an old inflection	• verb-noun idiomatic – combinations *kick the bucket*
III. COMPOUNDS		
• figurative compounds *lame duck* • semi-figurative compounds *high school* • functional compounds *police dog*	• semi-terms *magic eye*	• nominal compounds *golf club*, *connecting flight*
IV. PROVERBS		
• proverbs *too many cooks . . .* • quotations *to be or not to be* • greetings *good morning* • phatic phrases *have a nice day* • catch phrases *horses for courses*	• proverbs *half a loaf is better than no bread* • winged words *One small step for man . . .* • routine formulas *how do you do* • expletive constructions *give him an inch and . . .*	• sentence-like units *good-morning*
V. PHRASAL VERBS		
• phrasal verbs *get up*, *see trough*	• non/idiomatic particle verb *to run at/to bask in* • non/idiomatic reflexive verb *to enjoy yourself/to prostitute yourself*	• verb-particle constructions *take off* • prepositional verbs *refer to*
VI. LIGHT-VERB CONSTRUCTIONS		
• support verb constructions *to take a decision*	• noun phrase with semantically void verb *set in motion*	• light-verb constructions *to take a walk*
VII. PREPOSITIONAL PHRASES		
• compound prepositions *in spite of*	• MWEs with syntactic function *with regard to*	• prepositional phrases *in bed*, *in jail* • complex prepositions *on top of*

Kilgarriff and Rychlý 2010), enabling the identification and extraction of collocations from corpora. The Sketch Engine tool with its word sketches feature is frequently used (Kilgarriff et al. 2004), also for the compilation of collocation dictionaries. These mainly focus on the presentation of frequent, semantically relevant and syntactically regular word combinations.

In lexicography, FIXED PHRASES and IDIOMS are understood as MWEs that show at least some degree of fixedness in terms of structure and semantic idiomaticity. They can usually be found in separate sections in dictionaries, titled 'phrases' or 'idioms', combining several sub-types, which include proverbs and other sentence-like or pragmatic expressions, for example *hold your horses*, or *wild horses couldn't/wouldn't drag someone somewhere*.

In addition to idiomaticity, fixed phrases and idioms typically express metaphorical or pragmatic meaning, as their use depends on extralinguistic circumstances, for example *I could eat a horse* (expressing hunger). They can either show some lexical variation (as in *back/pick the right/wrong horse)*, or show lexical idiosyncrasy (as in *to and fro*), or else syntactic restrictions (e.g. *by and large*). Dictionaries sometimes provide detailed information about their use, and (more rarely) possible morphological or syntactic transformations or restrictions, but in general monolingual dictionaries limit the description to definitions and illustration of use. The two notable exceptions are SČFI (Čermák et al. 2009) and ODCIE (Cowie et al. 1983). The latter, especially vol. 2 on clause and phrase idioms, contains grammatical information and some information on syntactic variation. Explicit information is given for the commonest clause and phrase patterns by means of codes in square brackets and/or grammatical functions of collocates are identified. *Brook no delay*, for example, is followed by the code [V+O]; collocates functioning as the subject of *hold water* are listed after the code S: *theory, argument; explanation, reason, excuse; belief, need*. Some systematic structural changes are indicated: *give sb a good hiding*, for example, would be related to a structure with *get*, etc. Entries without explicit information are illustrated to cover for the most common variants of each grammatical type.

As opposed to lexicography, on the NLP side there were attempts to formalise the semantic composition of idioms and to make an association between the individual elements of the idiom and their (non-compositional) senses, such as *spill* in the sense of 'reveal', and *beans* as 'secret' in *spill the beans* (Baldwin and Kim 2010: 5). However, as MWEs vary in the degree of their decomposability, and some are in fact not decomposable at all, compositionality analysis would not be able to predict this regularity, as these senses for each word are not readily available outside a particular MWE.

Idiomatic COMPOUNDS are mostly defined as idiomatic (semantically non-transparent) word combinations functioning as different types of nominal phrases (e.g. *civil servant, connecting flight*, *old school*). They are included in dictionaries within a dedicated section of the entry or as headwords.[10] In terms of structure, they do not deviate from regular rules in morphology and syntax — they can take inflections, but they are fixed in form, as words typically cannot be added or removed from them. As such, they are difficult to separate from collocations and non-idiomatic compounds (e.g. *table leg*), which are spontaneously produced and found in their thousands in corpus data (Atkins and Rundell 2008: 169). From a lexicographic point of view, Atkins and Rundell recommend a three-point scale in determining the level of idiomaticity (mainly targeted at distinguishing between potential headwords and non-headwords): figurative compounds (*lame duck*), semi-figurative compounds (*high school*) and functional compounds (*house agent, police dog*). The first two represent potential headwords, but not the last one, as

functional compounds are very difficult to separate from productive, non-idiomatic compounds, like *house size* and *police pensions* (Atkins and Rundell 2008: 171).

From the NLP point of view, the property of noun compounds which has put them in the spotlight of NLP research is their underspecified semantics (Baldwin and Kim 2010: 12, Cordeiro et al. 2016, Schulte im Walde et al. 2016). Similar to lexicography, the NLP side also endeavours to find methods to identify compounds as explicit semantic representations in the mental lexicon (e.g. Schulte im Walde and Borgwaldt 2015). One popular approach to capturing the semantics of compound nouns is via a finite set of relations (e.g. *orange juice*, *steel bridge* and *paper hat* could all be analysed as belonging to the 'make' relation, where the head is 'made from' the modifier). This observation has led to the development of a wide range of semantic relation sets, based on abstract relations, direct paraphrases, such as by using prepositions or verbs, or various hybrids of the two. The downside of this approach is, according to Baldwin and Kim (2010: 13), in low inter-annotator agreement, poor coverage of data from different domains, and neglecting the impact of context on interpretation.

In relation to this type of MWEs, the NLP community has also shown significant interest in the syntactic disambiguation of MWEs with three or more elements or terms, as indicated by bracketing (Baldwin and Kim 2010: 13, 25):

glass window cleaner : (= [[glass window] cleaner]), (= [glass [window cleaner]]

According to Baldwin and Kim (2010: 25), the best-performing models for solving this task take features derived from both adjacency and dependency models, along with various syntactic and semantic features.[11]

In dictionaries, PROVERBS – as a separate MWE type – are usually found under 'phrases' or 'idioms' sections. Although several subtypes have been recognised (cf. Bergenholtz and Gouws 2014), they rarely receive an explicit meta-lexicographic label. From a lexicographic point of view they are interesting as a discrete category, since their meaning and use are fundamentally dependent on extralinguistic context. For example:[12]

- **don't change horses in midstream**
- *proverb* Choose a sensible moment to change your mind.
- **frighten the horses**
- *[usually with negative]* Do something likely to cause public outrage or offence.
- **from the horse's mouth**
- (of information) from the person directly concerned or another authoritative source.

Figuration is obviously considered an important feature in relation to this type of MWE (e.g. *take the bull by the horns*, *hold one's horses* etc.), similar to idioms and figurative compounds.

As shown in Table 1, so-called sentential MWEs are not in the centre of interest in NLP, although they are a legitimate part of the mental lexicon in relation to their semantic and pragmatic idiomaticity.

PHRASAL VERBS are MWEs consisting of a verb plus one or more particles (VPC)[13] that show a certain degree of idiomatic meaning. As such, they are of interest to lexicographers, who need to take into account their statistical, semantic, and syntactic properties to be able to identify them in text and handle them systematically in dictionaries.

a. VPCs are a frequent phenomenon in texts, but the distribution of their different semantic values is asymmetric. Some particles are frequently used with verbs (e.g.

combinations with *up, off* etc. in English) but not all combinations are idiomatic, and therefore not interesting for description in dictionaries (and not part of the mental lexicon).

b. VPCs show different degrees of semantic idiomaticity: they may have a literal meaning[14] and one or more figurative or metaphorical meanings. While some VPCs are clear candidates for lexicalisation in terms of their semantic idiomaticity (e.g. *make out*), others are semantically closer to the semantics of their component words (e.g. *check out, blow over*). In this respect, lexicographers find it difficult to determine even the basic form – if *get away* and *get away from* are one or two lexical units (Atkins and Rundell 2008: 173). Similarly, these characteristics present a problem in their identification and parsing from the NLP point of view.

LIGHT VERB CONSTRUCTIONS (LVC), combinations of light verbs or support verbs and nouns or prepositional phrases, such as *to give a lecture, to come into bloom*, are included in various parts of the dictionary macrostructure. According to Atkins and Rundell (2008: 173), it is typical of verbs in these constructions to carry less meaning than in many other contexts. However, in LVCs the degree of idiomaticity can also vary significantly. In many cases, the semantics of the verb are almost non-existent, and the LVC can be paraphrased with the verbal form of the noun complement (e.g. *take a walk* vs. *walk*), or in other cases the idiomaticity of the combination depends on the presence of a particular noun, for example *have a walk* vs **have a race*; *run a race* vs **run a walk*.

In terms of both morphology and syntax, LVCs can be unpredictable. As mentioned by Baldwin and Kim (2010: 16) LVCs can undergo passivisation (e.g. *an offer was made*), extraction (e.g. *How many offers did Kim make*?) and internal modification (e.g. *make an irresistible offer*).[15] On the other hand, there are hard constraints on what light verbs a given noun complement can be combined with (cf. **give/do/put/take an offer*). However, some noun complements can combine with multiple light verbs (e.g. *do/give a demo*), often with different semantics (e.g. *make a call* vs. *take a call* vs. *have a call*).

As mentioned above, LVCs can be found in various parts of the dictionary entry structure, either as lexical units, particularly in bilingual dictionaries, or as lexico-grammatical patterns under particular senses, especially within general monolingual dictionaries, for example:[16]

take (sense 21.) 'do or have something':

take a seat (= sit down)
take power/office
take cover (= hide)
take a risk
take (the) credit
take control/command
take responsibility

or even within an independent section (e.g. phrases), together with other types of MWEs.

PREPOSITIONAL PHRASES, which include categories such as 'compound prepositions' (Atkins and Rundell 2008: 169) or 'MWEs with syntactic functions' (Bergenholtz and Gouws 2014), are seldom included in dictionaries as independent and explicit MWEs,

and sometimes do not receive any lexicographic description. However, prepositional phrases with idiomatic meaning, such as *at hand, by air, in line*, can be found especially in EFL dictionaries due to their potential importance for achieving fluency in foreign language learning.

In contrast to being largely ignored by lexicographers, 'MWEs with syntactic functions' (e.g. *with regard to*) are interesting for NLP researchers as they can cause problems in parsing due to syntactic diversity, various degrees of markedness, productivity and modifiability, differing extents of semantic markedness, as well as high frequency. For these reasons, it is difficult to achieve the right balance between over- and under-sensitivity in the identification of this kind of MWEs.

3. Analysis of multiword expressions in selected lexicographic databases

In this section, we look at how a number of lexicographic and more NLP-oriented projects deal with MWEs in practice. The description of methods is based on an analysis of a set of samples which were provided in the context of the joint COST PARSEME/ENeL workshop on MWE e-lexicons (Skopje 2016). The workshop brought together 25 experts working on MWEs in lexicography and NLP. Prior to the workshop, participants were asked to provide a small sample of nouns, adjectives and verbs.[17]

In order to allow us to compare the data from the different projects, participants were asked to provide the following additional information for each sample:

1. information on the different MWEs types included in the project, and concepts adopted for the interpretation of MWEs;
2. information on the presentation of MWEs, particularly at the level of the inner structure of the MWEs;
3. information on the semantic information attached to MWEs (e.g. is a definition provided for MWEs and if so, for which types);
4. information on the lexicographic tools used in the project (e.g. Dictionary Writing System and Corpus Query System).

We received samples from the following projects:

1. The Algemeen Nederlands Woordenboek (ANW) for Dutch[18]
2. The Slovene Lexical Database (SLD)[19]
3. The multilingual Kamusi project[20]
4. Elhuyar bilingual dictionary[21]
5. Automatically extracted Noun-Verb expressions for Basque (Gurrutxaga and Alegria 2013)
6. Idion[22] for modern Greek
7. Polytropon[23] (Fotopoulou et al. 2014)

Below, we give a more detailed description of the projects. The description is followed by a summary of the main characteristics, i.e. the MWE types distinguished in the project, the internal structure of these MWEs and whether definitions are included.

Table 2. Main characteristics of MWE types in ANW

MWE types	collocations, *frequent free combinations*,[32] fixed expressions, proverbs, compounds
Phrase structure	collocations, for example A+N, V+N fixed expressions: variants of the canonical form
Definitions	fixed expressions, proverbs and hyperlinked compounds

The **ANW Project** is a freely-available online dictionary and lexicographic database of contemporary Dutch. Four types of MWEs are distinguished (Table 2): combinations (frequent free combinations and collocations), fixed expressions, proverbs, and compounds. All types are in separate sections and are quickly accessible from the dictionary interface. For combinations, the basic phrase structures are made explicit, for example A + N, V + N, N + PP. No inner structures are provided for the other types of MWEs, for example the fixed expressions, proverbs and compounds. Some variants of the canonical form are mentioned, such as variants with a diminutive and/or plural, as in *de kat/katjes in het donker knijpen* 'squeeze the cat/kittens in the dark', meaning 'be sneaky', or variants with inclusion or omission of the article, but there is no exhaustive description of all possible morphological variants of MWEs, like those with the verb in all conjugation forms. If relevant, fixed expressions are supplemented with semantic relations, for example synonyms or antonyms.

Definitions are provided for fixed expressions, proverbs and hyperlinked opaque compounds (which are defined in their own entries). For example, the compound *lapjeskat* 'calico cat' is defined in its own entry, whereas the semantically compositional combination *krolse kat* 'cat in heat' is listed in combinations with the structure A + N in the entry *kat*. Examples are given, but no definition, because the meaning can be reconstructed from those of *krols* and *kat*. The semantically non-compositional expression *cyperse kat* 'tabby', however, is listed under the fixed expressions of *kat*, because the expression does not have the meaning 'cat from Cyprus'. It is also given a definition and examples.

The Slovene Lexical Database (SLD) was compiled with two main purposes: to provide the basis for the compilation of various dictionaries of (modern) Slovene and to enable the development of applications and tools for Natural Language Processing of Slovene. It consists of 2,500 entries, with 2,053 multiword units, 1,500 phraseological units, and 49,000 collocations.

As the umbrella criteria for distinguishing different types of MWEs, the semantic properties of these expressions, such as semantic (non-)transparency or (non-) compositionality, and (non-)metaphorical meaning of the combination as a whole, are used. On this basis, on the first level collocations and extended collocations (i.e. collocations with an obligatory third element) and so-called syntactic combinations which do not need explanation, and therefore are semantically transparent as a combination, are distinguished. On the next level MWEs that need an explanation are registered. In general, two different entry elements are used (Table 3): (1) multiword units mainly covering compounds with nouns, adjectives and adverbs as headword components; and (2) phraseological units that cover idioms, proverbs and other multiword expressions that need explanation (catch phrases, winged words, greetings, etc.). Compounds, especially if they pertain to a particular domain, are briefly described with a label, sense indicator, or short gloss. Meaning and usage

of phraseological units are described with full sentence definitions, where all constituents of the broader context can be shown in a typical communicative pattern:

canonical form with lexical variants and obligatory open slots	*naliti/natočiti komu čistega vina* (lit. 'to pour pure wine to someone' = tell somebody an unpleasant truth)
label	mainly in political context
definition	*če rečemo, da smo komu nalili čistega vina, želimo poudariti, da smo mu povedali resnico ali dejstva o kaki neprijetni stvari ali dogodkih* (if we say that we poured pure wine to someone we want to emphasise that we told them the truth or facts about an unpleasant thing or event)
syntactic variants	*naliti si čistega vina / naliti komu čistega vina o čem / glede česa*
selected corpus examples	*Francija bo morala domači javnosti naliti čistega vina glede Evrope.*

Since SLD is primarily designed for lexicographic purposes, the inner syntactic structure of MWEs – except for collocations which are automatically extracted from the corpus – is not formalised.

The **Kamusi Project**[24] aims to produce a collaborative multilingual online dictionary. Included MWEs are understood as combinations of words with a unique meaning that cannot be derived from the sum of their parts. The dictionary, however, does not make distinctions between different types of MWEs (Table 4).

When a user adds a term to Kamusi, a script checks whether or not it is a multiword. Then the user specifies the language and the internal structure of the MWE by selecting search elements. The user then selects (1) whether the phrase is separable, and if so, where; (2) part of speech; and (3) attributes which are determined according to different parts of speech.

Finally, a definition is requested in the language of the entry, which can be translated to other languages. The definition is not a translation of the MWE, but a translated explanation of the meaning of the MWE. Users can also add examples, preferably by copy/pasting them from the source web site, along with the link.

The Elhuyar Foundation[25] provided resources that contain two types of MWE data: (1) MWEs within the entry structure in **Elhuyar bilingual dictionaries** (Table 5): Basque – Spanish, French and English; and (2) automatically extracted **Noun-Verb expressions** from the journal corpus (Gurrutxaga and Alegria 2013).

In the bilingual dictionaries MWEs are described in the subentry element according to the subentry type 'lexia' or 'lokuzioa'. In the first category noun phrases and nominal or

Table 3. Main characteristics of MWE types in SLD

MWE types	collocations, *extended collocations*, *syntactic combinations*, *phraseological units* (idioms, proverbs), compounds
Phrase structure	collocation, for example A+N, $N+N_{gen}$, $V+N_{acc}$ phraseological units: lexical variants and obligatory open slotssyntactic combinations: syntactic variants
Definitions	compounds, phraseological units

Table 4. Main characteristics of MWE types in Kamusi

MWE types	no distinction made (all MWEs are treated the same)
Phrase structure	lexicalised components with POS and attributes according to POSinformation whether the phrase is separable
Definitions	all MWEs

Table 5. Main characteristics of MWE types in Elhuyar bilingual dictionaries

MWE types	*noun phrases*, compounds, *conversational formulas*, idioms, collocations
Phrase structure	canonical form
Definitions	all MWEs

adjectival compounds are described (i.e. etxe-katu – 'domestic/house cat'). The second includes conversational formulas, idioms and collocations (i.e. lau katu – 'four cats'). For each MWE, the canonical form is registered and the connection to the lemma is indicated. The main focus is, however, on translation equivalents where the target language is specified, translations are provided for each sense of the MWE, usage types are specified and examples are translated.

In the second data set, automatically extracted N+V expressions from the journal corpus are presented in individual MWE elements (Table 6). For each N+V expression an MWE class is indicated, i.e. idiom, collocation, or free combination, the canonical form of the bigram is registered and morphological information about each component of the N+V expression, and the detected component variations. For each N+V expression, one example from the corpus is provided and the t-score, MI, and a few other association measures are included.

IDION[26] is an environment for discussing and encoding MWEs of Modern Greek. It describes verbal and nominal MWEs and no classification of types of MWEs is given (Table 7).

The project uses an xml-editor with a number of tabs on the users' interface. The Preview tab is auto-generated and offers an overview of the information about the entry. In the General tab, the meaning of the MWE is encoded. Exhaustive morphological and syntactic description of the MWE is encoded with the Forms tab. The encoding is theory-neutral and is aimed to serve as a basis for any type of parser. To this end, morphological tags are standardized (PAROLE).[27] The encoded syntactic relations are also kept theory-independent (and therefore minimal) and include (1) information about free constituents (Phrasal information for constituents that are realised with full phrasal structures + lexical information for constituents that are realised with weak pronouns); (2) binding and control relations; and (3) delineation of fixed/semi-fixed strings.

In the Diagnostics tab, the editor can indicate whether a verb MWE has a free subject or not, the number of constituents of the MWE, and whether the MWE passivizes or not.

For each MWE entry, a characteristic example, along with the phonetic transcription, PAROLE transliteration and English translation is provided at the Use tab. The Relations tab stores the semantic relations between MWEs, e.g. synonymy, opposites and verb alternations.

Polytropon[28] is a conceptual lexicon for Modern Greek and is intended for use in a number of NLP applications (Fotopoulou et al. 2014). It consists of 15,000 tokens and

Table 6. Main characteristics of MWE types in automatically extracted Noun-Verb expressions for Basque

MWE types	idiom, collocation, *free combination*
Phrase structure	canonical form (N+V)
Definitions	N/A

Table 7. Main characteristics of MWE types in IDION

MWE types	*verbal* and *nominal MWEs*
Phrase structure	For all MWEs: canonical form, component variants, information on external structure and syntactic alternations, information on fixedness, word order, morphological properties, shallow syntactic analysis and alternative forms
Definitions:	All MWEs

Table 8. Main characteristics of MWE types in Polytropon

MWE types	collocations, *fixed MWEs*
Phrase structure	syntactic pattern, variants
Definitions	fixed MWEs

includes single- and multi-word entries. MWEs are described on lexical, syntactic and semantic level. Lexical entry structure is divided into two blocks, in line with the Saussurian notion of sign: at SIGNIFIER level morphological, syntactic, and functional information about lemmas is encoded, and at the SIGNIFIED level word meanings are registered. Each MWE is registered in its inflected form (as string) and in its base form (as lemma). For each MWE, part-of-speech is indicated and MWE type, the main distinction being that between collocations and fixed MWEs (Table 8). Further morphosyntactic information is also encoded depending on the grammatical category of the MWE (i.e. declension of one or more constituents, only_singular or only_plural for nouns, etc.), and possible variants of the head verb of the MWE are listed. Collocations are further specified with respect to alternative lemmas. At the syntactic level, fixed elements and non-fixed elements of syntactic pattern and their features are formalized, and selection preferences are also applied to the arguments (such as +/-human). At the semantic level, semantic concepts are specified, a gloss and examples are provided, near synonyms (and antonyms) are grouped under the semantic class and register type can be specified.

4.1 Analysis of the datasets

To highlight key elements that could enhance the synergy between the lexicographic description of MWEs and needs of the NLP community, we focused on the MWE types included in the database, description of the inner structure of the MWE, and inclusion of semantic information. Table 9 summarises the results.

The databases include two essentially different **MWE types**, the first being frequent free combinations and collocations, identifiable in the corpus with an automated procedure

Table 9. Properties of MWEs in the dictionary and NLP databases

Project	MWE type	Inner structure				definition	example
		phras. struct.	variant	morphology of MWE element	contiguity		
ANW Dutch	Freq free comb. Collocations Fixed express Proverbs Compounds	for freq free comb, colloc	✓			✓ not for colloc, freq free comb	✓
SLD Slovene	Ext Collocations Collocations Syntactic combinations Compouds Phras. units	for colloc and syntac. comb	✓			✓ not for colloc and syntac. comb	✓
Kamusi multilingual	No distinction	✓	✓	✓	✓	✓	✓
Elhuyar (Basque) Bilingual dictionaries	Noun phrases Compounds Convrs. form. Idioms Collocations		✓				✓
Elhuyar (Basque) Noun-Verb expressions	Idioms Collocations Free comb.	✓	✓	✓			✓
IDION Modern Greek	Verb-MWEs Nominal MWEs	✓	✓	✓	✓	✓	✓
Polytropon Modern Greek	Collocations Fixed MWEs (diff. POS)	✓	✓	✓		✓	✓

using popular statistical measures in tools such as Sketch Engine (Kilgarriff et al. 2004). The second category are semantically non-decomposable MWEs, which require a description of their meaning as a whole. This category includes various MWE types described with different terminology or conceptual backgrounds (cf. Table 1). There are grey areas in the whole set, particularly between free combinations and collocations, as well as between collocations and compounds and fixed phrases. MWEs with syntactic variability (and different degrees of semantic idiomaticity), e.g. VPCs, LVCs and prepositional phrases, are not included in the selected datasets as separate categories. In terms of structure or syntax, these types are close to free combinations and have a vague semantic function. They are also quite productive in language and are therefore interesting for NLP. As such, they probably deserve more attention in lexicographic databases, especially as separate identifiable lexical units.

The databases differ in relation to how detailed the description of MWE **phrase structure** is, which mainly correlates with their primary intended use: for human users (ANW, SLD, Kamusi, Elhuyar bilingual dictionary) versus computational processing (IDION, Polytropon, Elhuyar Noun-Verb expressions). A good example of the synergy between the two is the collaborative Kamusi project, where the crowdsourcers themselves formalise the inner structure. MWEs such as free combinations and collocations typically include the formalisation of their element structure on the POS level, as they are automatically identified in the corpus based on that information, while MWEs with idiomatic and pragmatic meaning usually do not include a detailed description of the constituents. More attention is dedicated to these descriptions in NLP-oriented databases, in particular to morphological features, lexical variants, and the syntactic modification of constituents. The comparison clearly demonstrates that the various projects provide very different levels of description of the inner structure of MWEs, which indicates the need for a standardisation effort in relation to MWE description in dictionaries, lexical databases, and MWE lexicons to ensure better consistency and possible use of the data by the same NLP systems.

Semantic information is provided either in definitions, translation equivalents, or through semantic relations. Definitions are especially important in the case of MWE types with the lowest levels of semantic compositionality, e.g. fixed phrases, compounds and idioms. Collocations and frequent combinations do not include definitions, but they can be organised in semantic clusters (e.g. in SLD) or concepts (Polytropon,). Translation equivalents are crucial for multilingual databases (Kamusi, Elhuyar bilingual dictionary), and are of special importance for machine translation.

Some databases also include other types of information, such as statistical data (automatically extracted Noun-Verb expressions for Basque), phonetic description of examples (IDION), or user data from crowdsourcing (Kamusi, IDION). In terms of the **tools** used within the different projects, Sketch Engine's Word Sketch functionality is especially popular in lexicographically-oriented projects (ANW, SLD, cf. also Krek et al. 2015), while in other cases custom-made applications are used.

If we compare the information encoded in the different databases, we see that the classification of MWE types is not in the centre of interest in the projects which are mainly intended for NLP purposes. If the project is more lexicographically oriented, different MWE types tend to be specified (e.g. Elhuyar, Polytropon, SLD, and ANW). However, these projects use diverse terminology to describe the included MWE types, which indicates the need for a more standardised MWE classification, in particular to establish better synergy between lexicography and NLP.

4. Conclusion and prospects

From the methodological descriptions in the previous paragraphs we can draw some conclusions and formulate to-do-lists which could narrow the gap between lexicography and NLP. Overall, lexicographic resources (or at least the databases of dictionaries) should be structured as truly 'computerised dictionaries' (cf. Boguraev and Briscoe 1989) or lexical databases (cf. Fontenelle 1997) to enable optimal computer processing of these resources. On the other hand, NLP resources should be supplied with an interface to make them as 'human-readable' as possible. There is room for improvement in the interchangeability of existing materials, retrieval of new MWEs, the levels and granularity of descriptions, and the ease at which data can be transferred. These are discussed in more detail below.

Improvements in lexicography:

Enable easy **recognition** of MWEs in dictionary databases by categorising them explicitly, especially in cases when they are included in explanatory sections or as example sentences under single-word headwords. This enables extraction of all the MWEs from the dictionary database and matching example sentences that illustrate them, according to a standardised classification.

Develop **standards for canonical forms** of MWEs. For example, inclusion or omission of articles in MWEs beginning with a noun, use of possessives (*one's*, *someone's* or *somebody's*), use of to-infinitives or sentence-like expressions, omission of *to* (*kick the bucket*, *to kick the bucket*, *one kicks the bucket)*, etc.

Develop standards and find ways to **inventory variants** on the morphological, syntactic, lexical and semantic levels. Also allow for extension of binary structures; for human dictionary users and for NLP it is desirable to encode how binary strings combine to form longer strings (cf. extended collocations in SLD: collocations with an obligatory third element). Crowdsourcing methods can be helpful in this task, or a substantial amount of corpus examples in the databases for each MWE can be used to cover any variants. Coverage of verb forms and other morphological variants could be achieved through inclusion of more example sentences in the database.

Preferably, use theory-neutral **encoding of syntax and morphology**, which is suitable for different types of parsers and based on a standardized tagset. This enables a unified system of parsing MWEs and the examples that illustrate them.

Make selectional **preferences explicit** (e.g. +/- human), because this is useful for machine translation and in general for NLP tasks.

Improvements in NLP:

Include the MWEs that are encoded in dictionaries **in NLP-oriented lexicons**.

Develop tools which can **retrieve** frequent **MWEs**, relevant but infrequent MWEs, e.g. proverbs, non-contiguous MWEs and tools which retrieve less noise (from the lexicographers' point of view, for example, named entities are noise). Part of this task is also generating suggestions for canonical forms of MWEs.

Improve distributional methods to **distinguish** between **literal and figurative senses** of strings, based on the presence or absence of other words in a larger context.

Develop tools to **detect** morphosyntactic and lexical **flexibility**, word sense disambiguation and probability of idiomatic reading.

Allow for easy **transfer** from retrieval systems **to dictionary writing systems**, e.g. validation options for lexicographers in NLP-lexicons.

Most of the suggestions mentioned above have already been put into practice in many projects, but awareness of the mutual benefits that can be obtained from the lexicography and NLP communities working together still needs to be raised. To this end, it is important for both communities to provide openly available resources and methods to each other, and this can help foster a lasting cooperation to keep, share, and improve the best of both worlds.

5. Acknowledgement

The work described in this paper has been supported by the IC1207 PARSEME COST Action[29] and the IS1305 ENeL COST Action.[30]

Carla Parra Escartín is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 713567, and Science Foundation Ireland in the ADAPT Centre[31] (Grant 13/RC/2106) at Dublin City University.

Notes

1. Jackendoff (1997: 156) estimates that the number of MWEs in a speaker's lexicon is of the same order of magnitude as the number of single words.
2. In languages such as German, Dutch and Norwegian, the high productivity of compounds, without white space delimitation, means that at least in these languages single-word MWEs should be taken into consideration.
3. http://www.elexicography.eu/events/workshops/parsemeenel-skopje-2016/.
4. In Dutch, for example, the uninflected adjective in *een groot man* (a great man) has a meaning that is different from the meaning of the inflected adjective in *een grote man* (a tall man). Not many adjectives in Dutch have this markedness in the uninflected form, but it is still a case of idiomaticity at the morphological level.
5. Atkins and Rundell mostly quote English monolingual and bilingual dictionaries. In cases where examples were needed to illustrate some of the MWE types, we used similar resources that are freely accessible online.
6. A compositional meaning may not necessarily result from combining particular senses of individual parts. More precisely, we can say that the meaning of a phrase is compositional if it can be obtained in a regular way from the meaning of components and the syntactic structure of the phrase. This 'regular way' is rarely a sum, but most often a more complex operation (e.g. an intersection).
7. As Bergenholtz and Gouws point out, dictionaries with a focus on cognitive needs (i.e. dictionaries that provide additional data regarding language, culture or extralinguistic environment) need more accurate classifications. Based on this, they propose 20 different MWE categories (see Table 1).
8. Baldwin and Kim base their classification on the syntactic fixedness typology proposed by Sag et al. (2002), but add a morphosyntactic layer to it.
9. Cf. Evert (2005), Wiechmann (2008), and Pecina (2009), who discuss and review association measures in the domains of both lexical co-occurrence and lexicogrammatical co-occurrence.
10. This is especially true of monolingual learner's dictionaries, which tend to make all compounds into headwords (Atkins and Rundell 2008: 225).

11. For further details see Baldwin and Kim (2010: 25).
12. All examples are from English Oxford Living Dictionaries: https://en.oxforddictionaries.com/.
13. Phrasal verbs — often described also as verb particle constructions or particle verbs — are lexical units typical of and extensively described in English and other languages (particularly from the Germanic family). Among the languages included in the PARSEME project (Candito et al. 2016), those that did not include verb-particle constructions in their annotations are Brazilian Portuguese, Czech, French, Lithuanian, Maltese, Polish, Romanian and Spanish.
14. Although from the point of view of lexicographic description, definition is required only for idiomatic combinations, Atkins and Rundell (2008: 172) emphasise that for bilingual dictionaries literal uses are also interesting, since there may be a one-word equivalent in the target language.
15. Syntactic variants of LVCs can also be found in other languages. For examples see the PARSEME project, with the guidelines available at: http://parsemefr.lif.univ-mrs.fr/guidelines-hypertext/?page=060_Specific_tests_-_categorize_VMWEs/020_Light-verb_constructions.
16. MED: https://www.macmillandictionary.com/dictionary/british/take_1.
17. In this article only basic information and links to web pages are provided, if available. Technical details (availability, size etc.) can be consulted there.
18. http://anw.inl.nl/search.
19. http://eng.slovenscina.eu/spletni-slovar/leksikalna-baza.
20. https://kamusi.org/.
21. https://www.elhuyar.eus/en/site/community/nor-gara-en/fundazioa-en.
22. http://idion.ilsp.gr/.
23. http://goo.gl/SgQlxS at http://athena.clarin.gr/.
24. https://kamusi.org/.
25. https://www.elhuyar.eus/en/site/community/nor-gara-en/fundazioa-en.
26. http://idion.ilsp.gr/.
27. http://nlp.ilsp.gr/nlp/tagset_examples/tagset_en/.
28. http://hdl.grnet.gr/11500/ATHENA-0000-0000-2862-5.
29. http://www.parseme.eu
30. http://www.elexicography.eu
31. https://www.adaptcentre.ie/
32. Italic font indicates categories that are specific for a particular resource and are not included in the classification in Table 1.

References

A. Dictionaries

***Algemeen Nederlands Woordenboek*. http://anw.inl.nl. (ANW)**

Cowie, A. P., R. Mackin and I. R. McCaig. 1983. *The Oxford Dictionary of Current Idiomatic English*. Volume II: *Phrase, Clause and Sentence Idioms*. Oxford: Oxford University Press. (ODCIE)

Čermák, F. (Ed. in Chief) et al. 2009. *Slovník české frazeologie a idiomatiky I-IV*, Prague: Leda. (SČFI)

***Digitales Wörterbuch der deutschen Sprache*. https://www.dwds.de/. (DWDS)**

Macmillan Collocations Dictionary for Learners of English. Macmillan Publishers Limited, 2010. (MCD)

Oxford Collocations Dictionary for Students of English. Oxford: Oxford University Press, 2002. (OCD)

B. Other literature

Atkins, B. T. S. and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press.

Bahns, J. 1996. *Kollokationen als lexikographisches Problem. Eine Analyse allgemeiner und spezieller Lernerwörterbücher des Englischen. Lexicographica Series Maior 74*. Tübingen: Niemeyer.

Baldwin, T. and S. N. Kim. 2010. 'Multiword Expressions' In Indurkhya, N. and F. J. Damerau (eds), *Handbook of Natural Language Processing*, Second Edition, CRC Press, Boca Raton, USA, 267–292.

Benson, M. 1990. 'Collocations and General-purpose Dictionaries.' *International Journal of Lexicography* 3.1: 23–35.

Bergenholtz, H. and R. Gouws. 2014. 'A Lexicographical Perspective on the Classification of Multiword Combinations.' *International Journal of Lexicography* 27.1: 1–24.

Bergenholtz, H. and S. Tarp. 2003. 'Two Opposing Theories: On H.E. Wiegand's Recent Discovery of Lexicographic Functions.' *HERMES – Journal of Language and Communication in Business* 31: 171–196.

Boguraev, B. and T. Briscoe. 1989. *Computational Lexicography for Natural Language Processing*. London: Longman.

Bott, S. and S., Schulte im Walde. 2017. 'Factoring Ambiguity out of the Prediction of Compositionality for German Multi-Word Expressions.' In Markantonatou, S., C. Ramisch, A. Savary and V. Vincze (eds), *Proceedings of the 13th Workshop on Multi-Word Expressions*. Valencia, Spain, April 2017. Association for Computational Linguistics, 66–72.

Breidt, E., F. Segond and G. Valetto. 1996. 'Formal Description of Multi-Word Lexemes with the Finite-State Formalism IDAREX'. In Tsujii, J. (ed.), *Proceedings of the 16th conference on Computational linguistics - Volume 2 (COLING '96)*. Association for Computational Linguistics, 1036–1040.

Buendía Castro, M. and P. Faber. 2014. 'Collocation Dictionaries: A Comparative Analysis.' *Monografías de Traducción e Interpretación (MonTI)*, 6: 203–235.

Burger, H., A. Buhofer and S. Ambros. 1982. *Handbuch der Phraseologie*. Berlin, New York: de Gruyter.

Candito, M. et al. 2016: *Annotation Guidelines: PARSEME Shared Task on Automatic Identification of Verbal MWEs: Version 1.6b* (last updated on November 26, 2016. Accessed on 25th Jun 2017: http://parsemefr.lif.univ-mrs.fr/guidelines-hypertext/).

Constant, M. et al. 2017. 'Multiword Expression Processing: A Survey.' *Computational Linguistics* 43.4: 837–892.

Cordeiro, S., C. Ramisch, M. Idiart and A. Villavicencio. 2016. 'Predicting the Compositionality of Nominal Compounds: Giving Word Embeddings a Hard Time'. In Erk K. and N. A. Smith (eds), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 1986-1997.

Church, K. W. and P. Hanks. 1990. 'Word Association Norms, Mutual Information, and Lexicography'. *Computational Linguistics* 16.1: 22–29.

Evert, S. 2005. *The Statistics of Word Co-Occurrences: Word Pairs and Collocations*. Ph.D. thesis, Stuttgart: University of Stuttgart.

Fillmore C. J., P. Kay and M. C. O'Connor. 1988. 'Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone.' *Language* 64.3: 501–538. Linguistic Society of America.

Firth, J. R. 1957. *'Modes of Meaning.' Papers in Linguistics, 1934-1951*. Oxford: Oxford University Press.

Fontenelle, T. 1997. 'Using a Bilingual Dictionary to Create Semantic Networks.' *International Journal of Lexicography* 10.4: 275–303.

Fotopoulou, A., S. Markantonatou and V. Giouli. 2014. 'Encoding MWEs in a Conceptual Lexicon.' In: Kordoni, V., M. Egg, A. Savary, E. Wehrli and S. Evert (eds), *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*. Gothenburg, Sweden: European Association of Computational Linguistics (EACL), 43–47.

Fraser, B. 1970. 'Idioms within a Transformational Grammar.' *Foundations of Language*, 6.1: 22–42.

Geeraerts, D. 1997. *Diachronic Prototype Semantics: A Contribution to Historical Lexicology*. Oxford: Clarendon.

Gurrutxaga, A. and Y. Alegria 2013. 'Combining Different Features of Idiomaticity for the Automatic Classification of Noun+Verb Expressions in Basque.' In Kordoni, V., C. Ramisch and A. Villavicencio (eds), *Proceedings of the 9th workshop on Multiword Expressions (MWE 2013). Volume: 13*. Atlanta, Georgia: Association for Computational Linguistics, 116–125.

Hausmann, F. J. 1985. 'Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels.' In Bergenholtz, H. and J. Mugdan (eds), *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch, 28.–30.06.1984*. Tübingen: Niemeyer, 118–129.

Heid, U. 2004. 'On the Presentation of Collocations in Monolingual Dictionaries.' In Williams, G. and S. Vessier (eds), *Proceedings of the 11th EURALEX International Congress*, Vol. 2, No. 79, 729-738.

Hoey, M. 2005. *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Hunston, S. and G. Francis. 2000. *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Jackendoff, R. 1997. 'Twistin' the Night Away.' *Language* 73: 534–59.

Kallas, J., A. Kilgarriff, K. Koppel, E. Kudritski, M. Langemets, J. Michelfeit, M. Tuulik and Ü. Viks. 2015. 'Automatic Generation of the Estonian Collocations Dictionary Database'. In Kosem, I., M. Jakubiček, J. Kallas and S. Krek (eds), *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., 1-20.

Kilgarriff, A. and P. Rychlý. 2010. 'Semi-Automatic Dictionary Drafting.' In De Schryver Gilles-Maurice, (ed.), *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks*. Kampala: Menha Publishers, 299–312.

Kilgarriff, A., P. Rychlý, P. Smrz and D. Tugwell. 2004. 'The Sketch Engine.' In Williams, G. and S. Vessier (eds), *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6-10, 2004, Vol.1*. Lorient: Université De Bretagne Sud, 105-116.

Kozareva, Z. and E. Hovy. 2010. 'A Semi-Supervised Method to Learn and Construct Taxonomies Using the Web'. In Lapata M. and H., Tou Ng (eds), *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 1110–1118.

Krek, S., A. Abel and C. Tiberius. 2015. 'Survey – WG3 ENeL: Dictionary Writing Systems & Corpus Query Systems.' *COST ENeL WG3 meeting. Vienna, 12 February 2015*. http://www.elexicography.eu/wp-content/uploads/2015/04/ENeL_WG3_Vienna_DWS_CQS_final_web.pdf.

Manning, C. and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Mel'čuk, I. 1998. 'Collocations and Lexical Functions.' In Cowie, A. P. (ed.), *Phraseology. Theory, Analysis, and Applications*. Oxford: Clarendon Press, 23–53.

Michiels, A. and N. Dufour. 1998. 'DEFI, a Tool for Automatic Multi-Word Unit Recognition, Meaning Assignment and Translation Selection.' *Proceedings of the First International Conference on Language Resources & Evaluation*. Granada, Spain, 1179-1186.

Michiels, A. 1999. 'The DEFI Project'. *Informatique et Statistique dans les Sciences humaines XXXV*, 1.4: 110–114.

Michiels, A. 2000. 'New Developments in the DEFI Matcher'. *International Journal of Lexicography* 13.3: 151–167.

Moon, R. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: Clarendon Press.

Pecina, P. 2009. 'Lexical Association Measures and Collocation Extraction.' *Language Resources and Evaluation* 44.1-2, 137–158.

Ramisch, C. 2015. *'Multiword Expressions Acquisition: A Generic and Open Framework', Theory and Applications of Natural Language Processing Series, XIV*. Springer.

Ramisch, C. 2017. 'Putting the Horses Before the Cart: Identifying Multiword Expressions Before Translation.' *Computational and Corpus-Based Phraseology: Second International Conference, Europhras 2017 London, UK, November 13–14, 2017*, 69-84.

Sag, I., T. Baldwin, F. Bond, A. Copestake and D. Flickinger. 2002. 'Multiword Expressions: a Pain in the Neck for NLP.' In Gelbukh, A. (ed.), *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, 1-15.

Santus, E., A. Lenci, Q. Lu and S. Schulte im Walde. 2014. 'Chasing Hypernyms in Vector Spaces with Entropy.' In Shuly Winter, I., S. Goldwater and S. Riezler (eds), *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Gothenburg, Sweden: Association for Computational Linguistics, 38–42.

Savary, A., M. Sailer, Y. Parmentier, V. Rosén, A. Przepiórkowski, C. Krstev, V. Vincze, B. Wójtowicz, G. Smørdal Losnegaard, C. Parra Escartin, J. Waszcuk, M. Constant, P. Osenova, F. Sangati. 2015. 'PARSEME – PARSing and Multiword Expressions within a European Multilingual Network.'In Vetulani, Z. and J. Mariani (eds), *Proceedings of the 7th Language & Technology Conference (LTC 2015), 27-29 November 2015*. Poznań, Poland.

Segond, F. and E. Breidt. 1995. 'IDAREX: Formal Description of German and French Multi-Word Expressions with Finite State Technology.' *Technical Report MLTT22*, Rank Xerox Research Centre, Meylan, France.

Schubert, K. 2014. *The Design of Prototypes of a Dictionary of English Collocations*. Accessed on 25th June 2017 https://www.uni-hildesheim.de/iwist-cl/projects-students/Projekt-Schubert-Kim.pdf

Schulte im Walde, S. and S. Borgwaldt. 2015. 'Association Norms for German Noun Compounds and their Constituents.' *Behavior Research Methods* 47.4:1199–1221.

Schulte im Walde, S., A. Hätty and S. Bott. 2016. 'The Role of Modifier and Head Properties in Predicting the Compositionality of English and German Noun-Noun Compounds: A Vector-Space Perspective.' In Gardent, C., R. Bernardi and I. Titov (eds), *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics (*SEM)*. Berlin, Germany, 148–158.

Shwartz, V., Y. Goldberg and I. Dagan. 2016. 'Improving Hypernymy Detection with an Integrated Path-Based and Distributional Method.' In Erk, K. and N. A. Smith (eds),

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), Berlin, Germany, 2389–2398.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. (ed.) 1987. *Looking up: an Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. London and Glasgow: Collins.

Stubbs, M. 2007. 'Inferring Meaning: Text, Technology and Questions of Induction.' In Mehler A. and R. Köhler (eds), *Aspects of Automatic Text Analysis*. Berlin: Springer, 233–253.

Svensén B.. 2009. *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge: Cambridge University Press. xvi.

Swanepoel, P. 1994. 'Problems, Theories and Methodologies in Current Lexicographic Semantic Research.' In Martin W. et al. (eds), *Proceedings of the 6th EURALEX International Congress*. Amsterdam, 11-26.

Wiechmann, D. 2008. 'On the Computation of Collostruction Strength: Testing Measures of Association as Expressions of Lexical Bias.' *Corpus Linguistics and Linguistic Theory 42: 253-290. [CrossRef][Mismatch]*

Wiegand, H. E. 1984. 'On the Structure and Contents of a General Theory of Lexicography'. In Hartmann, R. R. K. (ed.), *LEXeter '83 Proceedings*. Tübingen: Max Niemeyer, 13–30.