Representing Multiword Expressions in Lexical and Terminological Resources: An Analysis for Natural Language Processing Purposes

Carla Parra Escartín,[1] Gyri Smørdal Losnegaard,[1]

Gunn Inger Lyse Samdal,[1] Pedro Patiño García[2]

[1]University of Bergen, Norway
[2]NHH Norwegian School of Economics, Norway
Carla.Parra@uib.no, Gyri.Losnegaard@uib.no, Gunn.Lyse@uib.no, Pedro.Patino@nhh.no

Abstract

In the context of standardisation and interoperability of Language Resources and Tools (LRT), this paper addresses the formal representation of multiword expressions (MWEs) for Natural Language Processing (NLP) purposes. By formal representation we mean the encoding of MWEs in lexical and terminological databases. The representation should render a language resource maximally reusable and ideally allow for seamless integration into any type of NLP application. In the case of MWEs, the situation is particularly complex due to their lexical properties on the one hand, and morphosyntactic variation on the other. Furthermore, their representation in multilingual resources poses even bigger challenges due to extensive translational asymmetry. In this paper we discuss the challenges posed by the formal representation of MWES. We analyse the needs of four different projects, all NLP oriented, but with slightly different approaches to the collection and representation of MWES. Based on the analysis, we identify a minimal set of features to be accounted for in any formal representation of MWES, as well as a set of more specific task-dependent requirements hinging on the intended use of the lexical resource. Finally, we assess to what extent existing standards meet these requirements.

Keywords: Multiword Expressions, Harmonisation, Standardisation, Interoperability, Natural Language Processing Applications, Terminological Resources, Language Resources

1. Introduction

Lexical Language Resources and Tools (LRT), such as machine-readable dictionaries and lexical and terminological databases, constitute a key element of advanced Natural Language Processing (NLP) systems. For the last two decades, researchers in computational lexicography have promoted the importance of designing a set of standards for the creation of reusable and interoperable lexical resources (Moreno Ortiz, 2000; Copestake et al., 2002; Francopoulo et al., 2006b; Francopoulo et al., 2009).

However, the lexis of a language is more than just single words, and in this regard there are still challenges to be overcome. Expressions such as "*fit as a fiddle*", "*give in*", "*pose a problem*" and "*as a matter of fact*" are multiword units that need to be appropriately represented in computational lexicons and yet are difficult to represent in a standardised manner. In their seminal “pain in the neck” article, Sag et al. (2001) point out that multiword expressions (MWEs) constitute a major bottleneck in NLP applications, and recent work and initiatives suggest that this is still the case[1]. Moon (1998), Sag et al. (2001) and Baldwin and Kim (2010) note that MWEs exceed word boundaries and have unpredictable properties. Research in the MWE field has also shown that one of the most salient and defining features of MWEs is their semantic non-transparency or non-compositionality. However, there is no widely agreed upon definition or typology of MWEs (Moon, 1998; Cowie, 1998; Sag et al., 2001; Baldwin and Kim, 2010, among others). We adopt a broad definition of MWEs as word combinations that form a unit at some level of linguistic analysis (Ramisch, 2012), and which deviate from regular language lexically, syntactically, semantically, pragmatically and/or statistically (Moon, 1998; Baldwin and Kim, 2010). Thus, although collocations are not always considered MWEs, we also include statistically marked or institutionalised collocations as a type of MWE. The aim of this paper is to capture all kinds of constructions that may pose problems in automatic analysis, and to determine which information should be recorded if such expressions are to be represented in a lexical inventory for NLP purposes. Managing to successfully represent MWEs in lexical and terminological resources is essential to ensure their successful integration in NLP applications, workflows and infrastructures.

The remainder of this paper will focus on this issue from different perspectives, based on four different use case scenarios. Particularly, we will concentrate on defining what information shall be recorded when including MWEs in lexical and terminological resources. How to encode such information will be the subject of further research.

In section 2, four different research projects dealing with MWEs are used as case studies, and their requirements as regards the representation of MWEs are discussed. Section 3 discusses how different standards may be used to formally represent MWEs, and the prerequisites needed to ensure that the final resource is reusable in NLP applications. Section 4 consolidates the results of our analyses and discusses the prerequisites for improved representations of MWEs, and sections 5 and 6 discuss future work to be carried out and sum up the main findings of the study reported here.

2. Case studies: Projects representing MWEs

In the creation of a new lexical or terminological resource the intended usage of such resource may condition its layout and the information recorded in it. In the case of resources including MWEs, what properties to record and represent will depend both

[1] http://multiword.sourceforge.net; http://typo.uni-konstanz.de/parseme

on the specific purpose and the type(s) of MWE. For certain purposes, a purely lexical account will do: if the end users of a MWE resource are human translators or second language learners, a simple entry with the MWE, its correspondence in the second language, and maybe examples of use, will be sufficient. However, if we intend to reuse the same resource within an NLP application, in order to ensure that the MWE is correctly processed, the computer will probably need additional information for each MWE unit, such as its morphosyntactic properties and its particular behaviour. Different kinds of MWEs may also have different intrinsic features, and the information needed for each particular entry will thus vary with the type in question as well as with the intended final usage of the resource.

In the following subsections four different research projects dealing with MWEs are presented. These projects have been selected because they have been or are currently being carried out by the authors and are presented in chronological order. Two of the projects approach MWEs from a mainly monolingual perspective (subsections 2.1 and 2.4). The other two are multilingual and concern translational correspondences (subsections 2.2 and 2.3).

Each subsection starts with a brief summary of the research project and then proceeds to briefly discuss what information should be recorded for each MWE in the frame of that particular project. Project-specific requirements are then discussed and analysed further in section 4, focusing on the properties that should be mandatory in the representation of MWEs.

2.1 Collocations and statistical analysis of n-grams: Multiword expressions in newspaper text

Lyse and Andersen (2012) describe an empirical study carried out in 2009 which applied various statistical association measures (AMs) to two- and three-word sequences (bigrams and trigrams) from the Norwegian Newspaper Corpus (NNC)[2]. The aim of this study was to determine which AMs are better at picking out relevant MWEs representing different lexical and terminological categories.

The NNC contains ca. 1.3 billion running words and it is the largest searchable corpus of contemporary Norwegian language. With such large amounts of data, efficient tools to identify different kinds of MWEs automatically are of great interest. In fact, recurring MWEs could be thus systematically identified, correctly segmented and added to lexical databases. This could in turn improve the syntactic tagging of the corpus since certain MWEs could be stopped from being further processed by the tagger. Moreover, technical terminology is often realised as MWEs, and the identification of recurrent collocational patterns is relevant for term extraction, even in non-technical texts such as newspaper language.

[2] http://avis.uib.no/om-aviskorpuset/english

Within the context of this study, nine common AMs were applied to bigrams in the NNC and four AMs to trigrams. To analyse the behaviour of each AM in more detail, the 500 top-ranked MWE candidates for each AM were classified manually. A relatively broad definition of MWEs was adopted, taking MWEs to be words that co-occur so often that they are perceived as a linguistic unit. The high-ranked terms were classified according to the following set of categories: *anglicism MWE*, *foreign MWE* (e.g. Latin expressions), *grammatical MWE* (e.g. multiword adverbs), *idiomatic phrase*, *term candidate* and *concept structure appositional phrase* (a term preceded by its superordinate concept).

Table 1 presents some examples of the kinds of MWEs that were highly ranked in the study.

In order to record the identified MWEs, the main requirement would be a standardised way of expressing statistical information about the rank of an item, preferably also including information about the raw frequencies on which the rank was based and the AM used. The extraction of n-grams and their statistical ranking in Lyse and Andersen (2012) did not rely on any linguistic annotation of the data, such as part of speech or lemma information.

The manually categorised MWE units could be interesting for reuse as a gold standard for new statistical experiments, which then imposes further formal representation requirements. To represent foreign MWEs, such as the anglicism "*consumer confidence*" and the Latin expression "*annus horribilis*", additional attributes for encoding the meaning of the expression itself and the language in which they appear would be needed. Furthermore, foreign expressions raise the need to emphasise that some expressions maintain a foreign inflectional paradigm (e.g. the anglicism "*practical joke*" *(sg.)*, "*practical jokes*" *(pl.)*) whereas others adopt the Norwegian one ("*walkie-talkie*" *(sg.)*, "*walkie-talkier*" *(pl.)*) and some are only used as frozen expressions without a productive inflectional paradigm ("*freezing fog*"). For term candidates, such as "*alternative energikilder*" (alternative energy sources), morphological information about inflection and internal structure is also necessary.

Multiword unit	English translation	Suggested classification
consumer confidence	-	anglicism MWE
annus horribilis	(Lat.) horrible year	foreign MWE
etter hvert	gradually	grammatical MWE
grøss og gru	shiver and horror	idiomatic phrase
alternative energikilder	alternative energy sources	term candidate

Table 1: Examples of high-ranked collocations in our study

2.2 English and Spanish specialised collocations found in Free Trade Agreements

This project is aimed at approaching the study of the type of collocations that appear in specialised texts from the subject field of international trade, i.e. legal and economics texts. The project also concerns the formal representation of these lexical units, in such a way that the data is machine readable and thus, interchangeable across different language resources (Litkowski, 2006). The data were obtained from the FTA parallel corpus (Patiño García, 2013), with English and Spanish data drawn from 16 official Free Trade Agreements (FTA) including texts from the American and European varieties of the two languages.

Within the frame of this project, a specialised collocation is defined as a type of MWE composed of at least one term that serves as the node. The collocates of this term can be nouns, verbs, adjectives or adverbs in a direct syntactic relation with the node and which do not necessarily appear adjacent to it.

Collocations constitute a challenge for several reasons. First, they can be unpredictable lexical combinations, appearing either adjacent to each other or in a span of several words to the left or right of the node word. Second, in a specialised context, terminology alone is not enough since it is also necessary to master the collocations that are used with these terms. Third, non-experts may encounter problems producing the correct verb, noun or adjective that is typically combined with a specific term (Bartsch, 2004; L'Homme, 2009). However, the lexical combinations of terms do not receive enough attention in lexicography and terminography and are therefore underrepresented in language resources (Pavel, 1993).

English	Spanish
accord favorable treatment	otorgar trato favorable
labor or environmental law enforcement	cumplimiento de la legislación laboral o ambiental
prescribe a conformity assessment procedure	exigir un procedimiento de evaluación de conformidad
prepare | adopt | apply a technical specification	preparar | adoptar | aplicar una especificación técnica

Table 2: Specialised collocations in English and their Spanish equivalents [Source: FTA Corpus]

Table 2 presents some English and Spanish examples of specialised collocations that appear in the FTA parallel corpus. In order to produce a language resource which is reusable and interoperable, particular features of every specialised collocation should be properly represented. First of all, the node of the collocation shall be properly detected and annotated as a term used in a specific subject field. Secondly, all collocates that this term may take should be appropriately tagged as well together with the subject field in which this collocation occurs. In addition to this, information on syntactic and morphological, as well as dialectal, aspects should be included to account for the multiple realisations of these collocations in different varieties of the same language.

2.3 Spanish MWEs as the translational correspondence of German compounds

This project deals with nominal compound words in German and their phraseological correspondences in Spanish. The project aims at improving 1:n word alignment within Germanic and Romance languages and the automatic extraction of compound dictionaries. Such dictionaries need to be appropriately encoded to ensure their reusability, and thus the question of how to represent the correspondence between one word in a language and an MWE in another arises.

Spanish translational correspondences of German compounds usually have the form of regular noun phrases. However, they need to be appropriately represented to yield satisfactory results in NLP applications such as Machine Translation (MT) systems and Terminology Extractors. As an illustration of the kind of units studied in this project, Table 3 shows some of the German compounds found in the TRIS corpus[3] (Parra Escartín, 2012) and their translations into Spanish.

German compound	Compound constituents	Spanish correspondence
Wohnungsförderungsverordnung	Wohnung·s·förderung ·s·verordnung	Ley de promoción de viviendas
Warmwasserbereitung	Warm·wasser· bereitung	preparación de agua caliente
Wärmepumpeanlagenförderung	Wärme·pumpe· anlagen·förderung	promoción de instalaciones de bombas de calor

Table 3: German compounds and their correspondences into Spanish
[English: *Housing Promotion Act / Water heating / Promotion of heat pumping systems*]
[Source: TRIS Corpus]

As can be observed in Table 3, German compounds constitute a single unit and thus their formal representation does not seem particularly problematic. However, their Spanish translational equivalents may indeed pose a challenge for bilingual and/or multilingual projects, as their representation will need to be more detailed and complex.

[3] The TRIS corpus has been compiled for the purposes of the project described here.

As far as German compounds are concerned, it would be desirable to have an indication as to which is the "head" of the compound as it selects inflection and gender. This is usually the most-right element of the compound. Moreover, additional morphological information as regards the rest of the elements forming part of the compound and their internal structure would also be desirable as this conditions the translation of a compound. For instance, the fact that the word "*Anlage*" appears in plural in the middle of the third compound ("*Wärmepumpe**anlagen**förderung*") requires the Spanish translation to be plural as well ("*instalaciones*") and translating it in singular would imply a semantic change.

It is also necessary to indicate which elements may be inflected in general language but are fixed or semi-fixed when part of the nominal phrase which translates into German as a compound. And finally, it would also be important to indicate whether other modifiers could be accepted (e.g. an adjective preceding the nominal compound in German) and their position within the nominal phrase in Spanish.

2.4 An NLP study of Norwegian MWEs

The last project is still at an initial stage. It aims to build the first extensive inventory of MWEs for Norwegian, which will serve as a basis for a typology of Norwegian MWEs and for the integration of different types of MWE into NorGram, a computational LFG grammar for Norwegian[4]. The representation requirements presented here are preliminary results based on a pilot analysis of MWE candidates identified during the annotation of the Norwegian treebank INESS[5]. The MWEs in Table 4 are taken from the first chapter of the novel *Sofies verden* (*Sophie's world*) by Jostein Gaarder. They exemplify, although not exhaustively, different kinds of MWEs found in this text.

Norwegian MWE	Literal translation	Idiomatic translation
snakke om	talk of, about	talk about
stå igjen	stand again	be left, remain
gjøre lekser	do homework	do (one's) homework
skille lag	divide team	split, part (ways)
komme rekende på en fjøl	come drifting on a board	come from nowhere (with origin unknown)
sikker på	sure on	sure that, sure of/about
et eller annet	one or other	something

Table 4: MWEs in *Sofies verden*

[4] http://iness.uib.no/redmine/projects/inesspublic/wiki/NorGram_documentation

[5] http://iness.uib.no

The verbal MWEs in Table 4 exemplify verb-preposition constructions ("*snakke om*"), verb-particle constructions ("*stå igjen*"), verb-object constructions ("*gjøre lekser*" and "*skille lag*"), and idioms ("*komme rekende på en fjøl*"). Each of these types of MWEs has different inherent features that need to be accounted for correspondingly. Verb-preposition and verb-particle constructions tend to be syntactically quite flexible, as opposed to idioms, for instance. On the other hand, we may have different degrees of semantic compositionality even within the same category. In the case of verb-object combinations, there may be expressions whose meaning is fairly transparent, such as the light-verb (or support verb) construction "*gjøre lekser*", while in other cases the meaning is contributed by all the component words and is less transparent, such as "*skille lag*" (lit. "divide team"). Last but not least, it is also important to highlight that idioms also pose challenges as regards their formal representation because they are syntactically restricted. In the more idiomatic of the two verb-object examples, "*skille lag*", the object "*lag*" cannot take a determiner and must be in singular and indefinite form. The idiom "*komme rekende på en fjøl*" cannot be passivised without losing its figurative meaning, the verb "*reke*" ("drift") must be in present participle form, and the object noun "*fjøl*" ("board") must be in singular indefinite form. It is semantically non-transparent, and like most idiomatic expressions, its lexical components and their morphological form are fairly invariable (Moon, 1998).

If we now focus on the non-verbal MWEs in Table 4, differences arise again with respect to syntactic flexibility and semantic transparency. "*Sikker på*" is an adjective-preposition construction which fills the same syntactic function in the sentence as a simple adjective. Like prepositional verbs, adjectives with selected prepositions require a clausal or nominal argument, and they are transparent in meaning. "*Et eller annet*" (literally "one or other") functions as a pronoun at clause level. Its meaning is semi-transparent, and it is syntactically fixed in the sense that the word order is invariable and no other words may intervene. However, the disjuncts "*et*" and "*annet*" inflect, and must agree in gender with its anaphoric referent.

The MWE candidates compiled in this project will be stored as entries in a database. For the most general level of use, each entry will contain lexical information as typically found in dictionaries, such as lexical category (part of speech), definition, canonical form (dictionary entry form), surface form (the instance as it occurs in the source text) and, if relevant, context (the sentence from which they were extracted). For research documentation and organisational purposes, it will be necessary to supply each MWE instance with a unique identifier and an identifier for the MWE "lemma". Information about the source (type, genre, publication date, author etc.), the method used to extract the MWE, the MWE frequency, and pointers to other occurrences of a given expression will also be recorded.

Further, to ensure an adequate level of description for an empirically based, formal classification of MWEs, it will be relevant to know on which linguistic level(s) the MWE exhibits anomalous behaviour, as well as its degree of semantic transparency and syntactic flexibility. As MWEs have varying degrees of semantic transparency and syntactic flexibility, they should be described with reference to a semantic scale ranging from totally transparent in meaning to completely opaque, and a syntactic scale ranging from syntactically flexible to completely restricted (or fixed). Finally, it will also be necessary to represent the internal structure and the morphosyntactic restrictions of each MWE, such as the argument structure of idioms. Whether the relevant properties for each MWE will be identified through manual analysis or by using automated methods is an open methodological question at this stage of the project. However, bearing in mind that the database will be integrated in a computational grammar, this information will have to be included in such a way that the resource can be easily integrated in the grammar and yet contain all relevant information for stand-alone usage.

3. Existing standards for representing MWEs

As we have shown in section 2, the formal representation of MWEs poses several challenges for resource developers, in particular if we aim at the interoperability and reusability of the lexical resource. From a monolingual perspective, a standard for formal representation will have to adequately account for the semantic and morphosyntactic properties of the overall expression and of the component words, internal structure and dependencies, syntactic variation, and potentially also regional language varieties for the given language. For instance, in Spanish, the English idiom "*it's raining cats and dogs*" may be "*está lloviendo a cántaros*" (lit. "it's raining pitchers"), "*caen chuzos de punta*" (lit. pointed "pikes are falling"), or "*llueven hasta maridos*" (lit. "it's raining husbands"), among others, depending on the regional variety of the speaker. For multilingual resources, translational correspondences must be accounted for, and the properties above must also be described for each language and/or language variety. If resource developers aim to create a scalable resource which can also be used by NLP applications, the formal representation of such a resource must also be compliant with the input format accepted by the tools that will process the resource.

Several projects have been undertaken in the last decades with the aim of unifying the coding of computational lexicons and terminologies through the creation of norms. The proposed standards are implemented by organisations, research groups, companies and professionals in the field and foster the exchange of information without losses or obstacles in transmission. Among these projects we can mention

GENELEX[6], MULTEXT[7], EAGLES[8], SIMPLE[9] and ISLE[10]. However, no standard has been broadly accepted thus far.

A quick look at the deliverables written in projects promoting the standardisation, interoperability and reusability of language resources (Rirdance and Vasiljevs, 2006; Hinrichs and Vogel, 2010; Calzolari et al., 2011; Monachini et al., 2011; Borin and Lindh, 2011) reveals that in the case of lexical and terminological resources, there are two standards that are commonly being used and fostered: TBX and LMF. Here, we also look at the TEI initiative, a well-known standard for general text encoding. Table 5 summarises the main features of the three standards.

Standard	**Monolingual**	**Bilingual**	**Encoding of morphosyntactic features**	
			MWE level	**Token level**
TBX	No	Yes	Yes	No
LMF	Yes	Yes	Yes	Yes
TEI	Yes	Yes	Yes	Yes

Table 5: Summary of standards and encoding

3.1 The TermBase eXchange format

If we first consider the TermBase eXchange format (TBX)[11], one of its main advantages is also one of its main drawbacks: its DTD is extremely flexible. This flexibility makes it possible for the user to customise the database and use attribute names suiting the project in which the termbase is created, but comes at the cost of interoperability since the resource will be incompatible with the representation requirements of NLP tools and applications. Furthermore, in TBX MWEs can only be registered as strings. Since they cannot be tagged in a fine-grained manner at token level, TBX prevents the possibility of processing non-fixed MWEs successfully with automatic methods. For instance, it would be impossible to account for the fact that in English the idiom "*it's raining cats and dogs*" may take internal modification as in "*it's* ***certainly*** *raining cats and dogs today*". Furthermore, although it would be possible to represent the MWE in all tenses (e.g. "*it* ***is/was/will be/has been*** *raining cats and dogs*") as separate entries, this is clearly not a very efficient way of dealing with its completely regular inflection. The TBX standard was created within the localisation industry and with translators and terminologists as its main target

[6] http://llc.oxfordjournals.org/cgi/content/abstract/9/1/47

[7] http://acl.ldc.upenn.edu/C/C94/C94-1097.pdf

[8] http://www.ilc.cnr.it/EAGLES/browse.html

[9] http://www.ub.edu/gilcub/SIMPLE/simple.html

[10] http://portal.acm.org/citation.cfm?doid=1118062

[11] http://www.gala-global.org/oscarStandards/tbx/tbx_oscar.pdf

users and it was primarily envisaged for the creation of bilingual and/or multilingual resources, not monolingual ones. Although it is not adequate for monolingual description, other important features such as the regional language variety and multilingual translational correspondences are easily encoded.

In short, in order for TBX to be appropriate for the encoding of MWEs, the names of attributes and values would need to be restricted and agreed upon. Granularity up to token level should be integrated as well as the possibility of assigning inflectional paradigms and other features to allow for language processing and generation in NLP applications. Finally, it should also allow for the proper representation of monolingual lexicons without requiring at least a second language. Until these requirements are met, TBX does not serve as an appropriate standard for encoding MWEs.

3.2 The Lexical Markup Framework

The Lexical Markup Framework is another of the standards encouraged by major standardisation initiatives. It was developed by the Technical Committee 37 of the International Organisation for Standardisation, Subcommittee 4 (ISO TC37/SC4[12]) and, as stated on their website[13], LMF was developed combining the best designs and methods from many NLP lexicons. However, it was developed for NLP use and not for human users, which is unfortunate since lexical resources are extremely useful in related fields such as second language acquisition. Among its features, there is an extension for bilingual or multilingual dictionaries, designed to express equivalence relations applicable in automatic translation (ISO, 2008). It also includes a module for the representation of MWEs, known as NLP Multiword Expression Pattern, which allows the representation of the internal structure of fixed, semi-fixed and flexible lexical units in a computational lexicon (Francopoulo et al., 2006a; Francopoulo et al., 2006b; Francopoulo et al., 2009).

More recently, UBY-LMF has been published. UBY is a large-scale lexical-semantic resource based on LMF and has been developed with the aim of interoperability and the smooth integration of resources (Gurevych et al., 2012). Despite capturing lexical information at a fine-grained level, using ISOcat data categories and being directly extensible by new languages and resources, this LMF-compliant model currently fails to offer an appropriate representation of MWEs. In fact, MWEs seem to have been overlooked by the developers of this model since they have rather focused on the standardisation of the semantic encoding of the entries of lexical semantic resources.

However, a priori, LMF seems a promising candidate for the encoding of MWEs.

[12] http://www.iso.org/iso/standards_development/technical_committees/list_of_iso_technical_committees/iso_technical_committee_participation.htm?commid=297592

[13] http://www.lexicalmarkupframework.org/

Spohr (2012, p. 25 ff.) explores this possibility and acknowledges that although it is feasible to represent MWEs in LMF, this has several drawbacks which he further discusses after demonstrating the representation of "*throw to the lion*". In future work we will try encoding samples from our case studies in this format to test to which degree it actually meets all the encoding requirements we have detected for the different projects accounted for in section 2. The findings of Spohr (2012), however, seem to suggest that although it may be possible to represent MWEs successfully, such a representation might not be the optimal one.

3.3 The Text Encoding Initiative

The Text Encoding Initiative (TEI) also has a specific module for encoding dictionaries. The TEI guidelines (Sperberg-McQueen and Burnard, 2009) explain how to appropriately encode all relevant information for each entry. Concretely, page 262 offers an example in which a compound is encoded as part of a larger lexical entry. TEI dictionaries allow for the encoding of multiple properties of relevance for NLP applications, such as part of speech, geographical area and etymological information, and also include the possibility of adding links and cross-references to other entries in the same resource. This makes TEI particularly interesting for the encoding of lexical and terminological resources, even though it seems to have been disregarded by major standardisation and infrastructure initiatives. The main drawback of TEI – besides the fact that it is not encouraged by the major standardisation initiatives – is that it is very flexible, which again introduces the possibility that different resource developers use different approaches for the encoding of their resources.

4. Prerequisites for improved representations of MWEs

In the following we merge the requirements we have identified in the four projects described in section 2, offering an overview of properties that we believe should be mandatory in the formal representation of MWEs, regardless of the standard used. The differences in the nature of our research projects make us think that we have covered most of the main possible usages a lexical resource could have in NLP applications. As has also been discussed in section 3, existing standards do not currently seem to be fully appropriate for the encoding of MWEs. Although further analysis is required, it seems reasonable to conclude that a set of required features for the representation of MWEs needs to be agreed upon and that standards should comply with successfully encoding all those features. Spohr (2008) divides his requirements for the model of a multifunctional electronic dictionary into the categories *detail of description*, *access and retrieval*, *consistency and integrity*, *specific users' needs* and *specific needs of NLP applications*. He observes that "*[o]ne of the most striking requirements, which can be directly derived from the above analysis, is the fact that the underlying formalism cannot be entirely unconstrained, but rather has to be strongly typed*". This leads Spohr to propose the OWL formalism[14] for representation, a formalism based on the Resource Description Framework (RDF). Although we have not gotten as far as Spohr and we do not attempt here to define which formalism is best for representing MWEs, we have devised a modular representation schema which we believe would meet the requirements we identified. This schema, which has been designed after the representation model envisaged by META-SHARE, consists of three levels of detailed representation, one mandatory and two optional but recommended. We further suggest a need for optional type and purpose dependent representation schemas, or *encoding modules*. In the general, main schema (or module) described below, levels 1 and 2 both describe properties relevant for the description of the overall expression (type level). The second level is an extension of the first and targets more advanced users and usages, while the third level provides information about the MWE at token level. Ideally, levels 1 and 3 should be mandatory, but it is not feasible that every resource creator will be able to encode a potentially large number of expressions in such detail. We therefore propose level 1 as the minimum representation schema for every MWE, and thus the only mandatory level.

1. Type level (mandatory)
    a. Part of Speech (PoS)
    b. PoS standard
    c. Meaning
    d. The number of component words
2. Type level, extended description (optional)
    a. Canonical (base) form
    b. Level(s) of idiosyncrasy
    c. Translational correspondences
    d. Language variety
3. Token level (optional)
    a. PoS
    b. Lemma
    c. Grammatical features

4.1 Level 1: Type level

Many MWEs correspond syntactically to simple words or constituents in a sentence, such as the complex adverb "*etter hvert*" (lit. after each, "gradually") and the noun phrase "*preparación de agua caliente*" (lit. preparation of water hot, "water heating"). For such MWEs, the lexical category (part of speech, PoS) should be assigned (1a in the proposed schema). Not all MWEs correspond to one word or constituent, as is the

[14] Web Ontlolgy Language, http://www.w3.org/TR/owl- ref/

case with most verbal expressions. The specialised collocation "*accord favorable treatment*" in Table 2 and the verb-object construction "*skille lag*" (lit. divide team, "part") in Table 4 both exceed constituent level. It should thus be possible to express that the PoS category is "non-applicable". In such cases, the additional classification module could be used to assign the MWE a type label instead, such as *sentence* (like "*it's raining cats and dogs*"), *verb-particle construction* (VPC), *light verb construction* (LVC), etc.

Which PoS standard is used should also be accounted for (1b). Even though there is no specific standard that is commonly used in NLP, the PoS inventory (for European languages) normally includes the traditional categories noun, verb, adjective, adverb, pronoun, conjunction, preposition and interjection. Most linguists would probably not settle for such a crude classification, and for encoding purposes we recommend that the representation schema is equipped with the most widely used PoS standards. In case these are not applicable, the representation schema should also allow users to define their own custom-made inventories of lexical categories that are suitable for their individual projects or needs.

Meaning can be represented with a synonym, a definition, a translation or a transliteration. All of these possibilities should be available in the encoding schema (1c). 1d accounts for the number of constituent words.

All features at this level are mandatory, and features that are not relevant for a given MWE should be marked as non-applicable.

4.2 Level 2: Type level, extended description

Level 2 of our proposed schema targets more advanced usages and is recommended, but optional. After all, a particular resource may not be bilingual or account for dialectal varieties; or the MWEs may not have been analysed and thus may not be classified or described in terms of idiosyncrasy (at which linguistic levels they deviate from "regular" language; syntactic, semantic etc.). However, having a pre-defined module that envisages the addition of such information would ease the scalability and reusability of the resource in the long run. As for the canonical form, it would be desirable to have a standardised way of representing this, e.g. the base form of each component word.

4.3 Level 3: Token level

The final level describes the properties of the component words and again is recommended, but optional. This level allows for the annotation of component words with grammatical information.

4.4 Additional encoding modules

The provision of additional modules to the main schema will allow for optional representation of different types of MWEs, of information particular to a given field, topic or discipline, and of purpose-dependent properties. A modular representation schema thus makes it possible to describe MWEs from different perspectives according to the needs of the individual user or resource developer. Furthermore, optional modules for specialised information may simply be ignored by processing tools which do not make use of that particular type of information. Additional modules depending on the particular research project and the final usage of the resource could be:

- Classification
- Morphosyntactic profile
- Metadata
- Organisational data
- Semantic profile
- Terminology
- Multilinguality
- Named Entity

Due to the lack of agreement with respect to the definition and classification of MWEs, information about the type of MWE could be represented in a dedicated *classification* encoding module. This module should offer predefined MWE categories from existing typologies. It should also allow for customisation of classification schemas, so that users may classify the MWEs according to his/her own schema, and if desired, according to several schemas. Categories that reflect syntactic structure, such as *light verb construction* and *particle verb*, could be represented here, as well as more general types such as *collocation, idiom* or *metaphor*.

The description of the more complex morphological syntactic properties of an MWE would be difficult to account for at token level, since such properties often involve dependencies between words. We thus propose to have a dedicated *morphosyntactic* module. This would be the most important component for ensuring interoperability with and integration in NLP applications. The module should account for aspects that cannot easily be represented at word level, such as the internal structure of the MWE, morphosyntactic restrictions (e.g. the indication of morphosyntactically “frozen” words), subcategorisation information, description of internal modifiers, their type and position within the expression, etc. Dependency descriptions involve marking phrasal heads, node words and collocates, indicating which words take modifiers, etc. The module should also indicate the degree of syntactic flexibility, from fixed to completely flexible.

A *metadata* module would meet the requirements identified in 2.4, allowing for a description of the source material. This could be information about the source type (corpus, dictionary, website, etc.) and specific texts (title, author, date, etc.), and is particularly relevant for projects where MWEs have been extracted from multiple sources. The requirements pointed out in 2.1 further raise a need for *organisational data* such as the extraction method used, frequency and rank (based on the number of occurrences of the MWE in the source material), and pointers to other occurrences or entries.

The *semantic* module would be relevant for language analysis. This module should allow for an elaboration of the definition and meaning, the degree of semantic transparency, to which degree the different constituents contribute meaning to the overall expression, etc. Features relevant to terminology and multilingual resources are described in sections 2.1, 2.2 and 2.3 and include the representation of collocational features, ontological relations, etc.

5. Discussion and future work

The implementation of a flexible but standardised and agreed-upon encoding schema such as the one discussed here would ensure the scalability of lexical and terminological resources, since researchers could then take as a starting point an already developed resource and add the modules they need for their particular projects. For instance, the terminology resource described in 2.1 could be taken as a starting point for the creation of the resource under development in the project described in 2.4. Resources developed independently in different projects could easily be merged into one resource with several modules, where different modules encode the specific information for each project. Finally, in order to ensure the scalability and interoperability of the resources created, feature names, values and formats should be standardised to the extent possible and correspondences between different standards should be provided to ensure the successful merging of resources if necessary.

As a follow-up of the analysis reported here, we intend to assess the appropriateness of the different standards available for the encoding of lexicons and terminological databases, using data from our respective research projects. We may then determine to what extent these standards actually allow for encoding of the features that we have proposed as the minimal set of features to be included in the representation of any type of MWE and any type of NLP application. If we aim to develop resources which are standardised and interoperable, encoding MWEs in one of the existing standards would not be enough as it would be possible to have four different resources encoded using the same standard but providing different information or information with mismatched attribute names. In order to ensure the reusability of our resources, a compromise among all stakeholders is necessary by agreeing upon a standard set of attributes and values. This would make the mapping between different encoding formats feasible and as a result, merging, exchanging and enlarging resources would no longer be so problematic.

6. Conclusion

In this paper we have discussed the requirements for the formal representation of MWEs from different perspectives. Four projects have been presented, and their needs have been discussed to show the wide variety of projects and usage scenarios where an appropriate formal representation of MWEs may be relevant.

Despite several recent standardisation efforts and initiatives, none of the major encoding standards meet all of the requirements identified in section 2. In order to encode MWEs in lexical resources in a way that both accommodates our individual requirements and renders the resources comparable, extendable and applicable outside our own limited projects, we have thus proposed a modularised representation schema with different modules or profiles for different purposes and uses.

Importantly, this is not an attempt to define a new encoding standard. Rather, and as pointed out in section 4, we think that it is necessary to have more information considered as *mandatory* in the representation of MWEs, in particular with respect to the multilingual aspect and their unique features.

The projects described in section 2 present real usage scenarios, all of which require detailed formal representations of MWEs. From our point of view, efforts should be put into enhancing existing standards by devising DTDs with standardised sets of attributes and values for general descriptions and standardised ways for representing complex morphosyntactic information. The study reported here has highlighted the need for flexibility in the encoding of linguistic phenomena. Our recommendation is to implement specific modules for gathering and representing the specific information particular to a given topic, type or use for every MWE included within lexical or terminological resources. This will ensure their reusability and interoperability and will thus bring us closer to a proper treatment in NLP applications.

7. Acknowledgements

The research reported on in this paper has received funding from the EU under FP7, Marie Curie Actions, SP3 People ITN, grant agreement 238405, (project CLARA), and from the ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, grant agreement no. 270899, the University of Bergen and the Norwegian School of Economics.

The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

8. References

Atkins, S., Bel, N., Bouillon, P., Charoenporn, T., Gibbon, D., Grishman, R., Huan, C.-R., Kawtrakul, A., Ide, N., Lee, H.-Y., Li, P. J. K., McNaught, J., Odijk, J., Palmer, M., Quochi, V., Reeves, R., Sharma, D. M., Sornlertlamvanich, V., Tokunaga, T., Thurmair, G., Villegas, M., Zampolli, A., and Zeiton, E. (2001). Standards and Best Practice for Multiligual Computational Lexicons. MILE (the Multilingual ISLE Lexical Entry) Deliverable D2.2-D3.2. ISLE project: ISLE Computational Lexicon Working Group.

Baldwin, T. and Km, S. N. (2010). Multiword Expressions. In Indurkhya, N. and Damerau, F. J., editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

Bartsch, S. (2004). *Structural and functional properties of collocations in English: a corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Gunter Narr Verlag, Tübingen.

Borin, L. and Lindh, J. (2011). Deliverable D4.1: Metadata descriptions and other interoperability standards. Version 1.0, 2011-05-02. Deliverable in the META-NORD project (CIP 270899).

Calzolari, N., Bel, N., Choukri, K., Monachini, M., Odijk, J., Piperidis, S., Quochi, V., and Soria, C. (2011). Final FLaReNet Deliverable: Language Resources for the Future - The Future of Language Resources. The Strategic Language Resource Agenda. FLaReNet project.

Copestake, A., Lambeau, F., Villavicencio, A., Bond, F., Baldwin, T., Sag, I. A., and Flickinger, D. (2002). Multi- word Expressions: Linguistic Precision and Reusability. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands, pp*. 1941–1947.

Cowie, A. P. (1998). *Phraseology: Theory, Analysis, and Applications: Theory, Analysis, and Applications*. Clarendon Press.

Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., and Soria, C. (2006a). Lexical Markup Framework (LMF) for NLP Multilingual Resources. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, pp. 1–8, Sydney, Australia. Association for Computational Linguistics.

Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., and Soria, C. (2009). Multilingual resources for NLP in the Lexical Markup Framework (LMF). *Language Resources and Evaluation*, 43:57–70. 10.1007/s10579-008-9077-5.

Francopoulo, G., Declerck, T., Monachini, M., and Romary, L. (2006b). The relevance of standards for research infrastructures. In *International Conference on*

Language Resources and Evaluation - LREC 2006, Gênes/Italie. European Language Resources Association (ELRA).

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). UBY - A Large-Scale Unified Lexical-Semantic Resource Based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pp. 580–590, Avignon, France.

Hinrichs, E. and Vogel, I. (2010). Deliverable D5C-3: Interoperability and Standards. CLARIN Project.

ISO (2008). Language resource management - Lexical Markup Framework (LMF), ISO 24613:2008, ISO/TC 37/SC 4 N453 (N330 Rev.16).

L'Homme, M. C. (2009). A methodology for describing collocations in a specialised dictionary. In *Lexicography in the 21st century*, pp. 237–256. John Benjamins, Amsterdam/Philadelphia.

Litkowski, K. (2005). Computational Lexicons and Dictionaries. In Brown, K., editor, *Encyclopedia of Language and Linguistics (2nd ed.)*, pp. 753–759. Elsevier, London.

Lyse, G. I. and Andersen, G. (2012). Collocations and statistical analysis of n-grams - Multiword expressions in newspaper text. *Exploring Newspaper Language. Using the web to create and investigate a large corpus of modern Norwegian.*

Monachini, M., Quochi, V., Calzolari, N., Bel, N., Budin, G., Caselli, T., Choukri, K., Francopoulo, G., Hinrichs, E., Krauwer, S., Lemnitzer, L., Mariani, J., Odijk, J., Piperidis, S., Przepiorkowski, A., Romary, L., Schmidt, H., Uszkoreit, H., and Wittenburg, P. (2011). The Standards' Landscape Towards an Interoperability Framework. FLaReNet project.

Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford University Press.Moreno Ortiz, A. (2000). Diseño e implementación de un lexicón computacional para lexicografía y traducción automática. *Estudios de Lingüística del Español*, 9.

Parra Escartín, C. (2012). Design and compilation of a specialized Spanish-German parallel corpus. In Calzolari, N. C. C., Choukri, K., Declerck, T., Uğur Doğan, M., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Patiño García, P. (2013). *FTA Corpus: a parallel corpus of English and Spanish Free Trade Agreements for the study of specialized collocations*, volume 3 of *Bergen Language and Linguistic Studies*, pp. 81–92. University of Bergen Library, Bergen, Norway.

Pavel, S. (1993). Neology and phraseology as terminology-in-the-making. In

Terminology: applications in interdisciplinary communication, pp. 21–34. John Benjamins, Amsterdam.

Ramisch, C. (2012). *A generic and open framework for multiword expressions treatment: from acquisition to applications*. PhD thesis, University of Grenoble (France) and Federal University of Rio Grande do Sul (Brazil), Grenoble, France.

Rirdance, S. and Vasiljevs, A. e. (2006). Towards Consolidation of European Terminology Resources. experience and Recommendations from EuroTermBank Project. Technical report, EuroTermBank Consortium.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2001). Multiword Expressions: A Pain in hte Neck for NLP. In *In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pp. 1–15.

Sperberg McQueen, M. and Burnard, L. (2009). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Technical report, The TEI Consortium.

Spohr, D. (2008). Requirements for the Design of Electronic Dictionaries and a Proposal for their Formalism. In *Proceedings of the EURALEX International Congress 2008*.

Spohr, D. (2012). *Towards a multifunctional lexical resource design and implementations of a graph-based lexicon model*, volume 141 of *Lexicographica. Series maior*. De Gruyter.