Lexicographic Criteria for Selecting Multiword Units for MT Lexicons

Jack Halpern[1]

The CJK Dictionary Institute, Inc.
34-14-2, Tohoku, Niiza-shi, Saitama, Japan
jack@cjki.org

Abstract. A basic assumption in bilingual lexicography and machine translation (MT) is that the linguistic units of one language correspond to those of another language. But even in close language pairs, such as Spanish and English, there are numerous exceptions, while in some language pairs, such as English and Japanese, cross-linguistic lexical anisomorphism is so great that it becomes literally impossible to map certain words and phrases across these languages. This is especially true of linguistic units that consists of multiple components, or *multiword units* (MWUs). The recognition and accurate translation of MWUs play a critical role in enhancing the quality of machine translation[9]. In spite of the recent advances in MT based on neural networks (NMT), MWUs still present major challenges to MT technology. This paper discusses the fundamental principles for identifying and selecting MWUs for inclusion in bilingual dictionaries, both for humans and for MT systems (MT lexicons). It attempts to define the various subtypes of MWU based on lexicographic principles derived from extensive experience in bilingual lexicography, especially the compilation of a large-scale full-form lexicon for Spanish-English MT. It also introduces some large-scale resources designed to significantly enhance the translation accuracy of multiword proper nouns.

1 Introduction

1.1 The problem

The fundamental principle of selecting headwords in bilingual dictionaries is that *the words and phrases of one language can be mapped to those of another*. This is mostly true, but even in such relatively close language pairs as Spanish-English there are numerous exceptions. In some language pairs, such as English-Japanese or English-Chinese, lexical anisomorphism is so great that the principle of cross-lingual word/phrase correspondence often breaks down completely. That is, it becomes literally impossible to directly map certain linguistic units to those of the other language.

A string of words can be segmented into components in multiple ways. Linguists may disagree on how to combine these components to form meaningful linguistic units. The challenge is to decide which combination of components

qualifies as a lexical unit or dictionary headword. This paper attempts to define the various linguistic units that fall under the broad category of MWUs. These definitions are based on decades of experience in CJK lexicography, and on the compilation of a large-scale full-form Spanish-English lexicon (tens of millions of entries) for the Context-Based Machine Translation project[4] headed by Dr. Jaime Carbonell (MT expert and founder of the Language Technologies Institute at CMU).

1.2 Terminology

This paper defines and illustrates MWUs and their five subtypes, describing the criteria that qualify an MWU for inclusion as an entry in bilingual dictionaries. Intentionally, the terms "word" and "phrase" are avoided as formal categories because of their inherent ambiguity[7]. "Phrase," which can loosely refer to MWUs, is ambiguous and causes much confusion in language technology. It is often used in the sense of "any sequence of two or more words," or loosely in the sense of "compound word," without defining the relation between the components. However, for the sake of brevity, "word" below is used in the sense of *orthographic word*, defined as "an uninterrupted string of letters which is preceded by a blank space and followed either by a blank space or a punctuation mark"[10]. Two terms play a major role in defining MWU subtypes:

Lexical unit (or *lexical item*) is single free form ("word") or meaningful sequence of free forms or bound forms ("word elements") that constitute the basic elements of a language's lexicon. It is a distinctive unit of vocabulary that associates meaning with form. It is what the native speaker stores, or potentially stores, as a "word" or "phrase" in her internal lexicon, e.g. *house, in other words, take off, rain cats and dogs, unmarried, high school, headwaiter*. The near-synonym lexeme emphasizes all the members of an inflectional paradigm, rather than a specific wordform.

Lexical status refers to whether an MWU is a *meaningful* lexical unit (has a high degree of lexicalization); that is, whether it is (potentially) present in the internal lexicon of native speakers and functions as a meaningful syntactical/grammatical unit. On the whole, native speaker's intuitively feel that it is "a word or phrase" of their language. Thus *high school*, which is fully lexicalized, has lexical status, but *high building*, a free combination of words, does not.

1.3 MWU Subtypes

A **multiword unit** (MWU) is a combination of two or more words that commonly occur together. They may or may not function as a lexical unit, may or may not be semantically compositional, and may or may not have lexical status. This paper defines and analyzes five subtypes of MWUs.

Multiword expression (MWE): a lexical unit consisting of two or more words that together function as a single lexical unit.

Free word combination (FWC): a meaningful free sequence of words that follow the rules of syntax but has no lexical status.

Phraset: a recurrent meaningful free combination of words that has no lexical status in the source language but corresponds to a lexical unit in the target language.

Collocation: a recurrent combination of words co-occurring more often than by chance whose meaning is (mostly) compositional and transparent.

Multiword proper noun: a combination of two or more words that together function as a single proper noun.

Although the terms defined herein are based on morphological and lexicographic considerations, different linguists use these terms in somewhat different ways. It should be noted that the subtype categories defined, by their nature, are not necessarily rigorous, nor are they necessarily mutually exclusive.

2 Multiword Expressions

2.1 Definition

A **multiword expression** (MWE) is defined by linguists in different ways. Calzolari et al.[3] gives a general definition as "a sequence of words that acts as a single unit at some level of linguistic analysis." In *Introduction to the special issue on multiword expressions*, Villavicencio et al.[10] define it as "an expression for which the syntactic or semantic properties of the whole expression cannot be derived from its parts," while Sag et al.[8] define it "very roughly" as "idiosyncratic interpretations that cross word boundaries (or spaces)." Here it is defined as "a lexical unit consisting of two or more simplex words that together function as a single meaningful lexical unit."

Table 1. Additional characteristics of MWEs.

a) Zona residencial	residential zone (transparent compositional compound)
b) Dar a	look out onto (opaque non-compositional phrasal verb)
c) Elefante blanco	look out onto (opaque non-compositional phrasal verb)
d) Devanarse los sesos	rack one's brains over (idiomatic expression)
e) Matar dos pájaros de un tiro	kill two birds with one stone (opaque compositional proverb)
f) Lo antes posible	as soon as possible (locution)

MWEs have some additional characteristics: (a) they represent both content words and function words, (b) they have full lexical and lexicographic status, (c) some are monolingually compositional but bilingually non-compositional, (d) they can range from semantically transparent to opaque, and (e) they have high semantic cohesiveness. Some examples (see Table 1).

2.2 Analysis

It is important to understand MWEs with some precision, and to distinguish them from FWCs and phrasets, as difficult as this may be in the case of borderline cases. MWEs are groups of words that co-occur more frequently than by chance, have a high semantic cohesiveness (attraction between components) and, *most importantly*, represent a concept, often a well established designatum. They are the core backbone of a language, what native speakers intuitively feel are "the words and phrases" of their language.

2.3 Typology

MWEs can be classified into eight (or more) subtypes. Though on the whole the subtypes are mutually exclusive, there is some overlap between them.

Compound words are combinations of two or more words (free morphemes) or word elements (bound morphemes) that together function as single lexical item, usually transparent, like *learner's dictionary* (noun compound) or *take into account* (verbal compound). If they are opaque, they are normally called idioms. Some examples (see Table 2):

Table 2. Compound words.

Tinta china	india ink
Parada general	general strike
Lobo marino	sea lion
Caja fuerte	safe, strong box
Casa de campo	field house
Papel cuadriculado	graph paper

Phrasal verbs (*or verb particle constructions*) consist of a verb followed by one or more particles that together function as a lexical unit[10]. Some, like *acabar de* 'just' are idiomatic and opaque, while others have some transparent senses and some opaque or semi-opaque senses. For example, *fight on* 'continue to fight' is perhaps semi-opaque, but in the sense of 'fight on the top of', as in they *fought on the roof*, it is completely transparent. *Estar por* is opaque in the sense of 'be on the verge of' but transparent (compositional) in the sense of 'be for'.

Idioms, and other **lexicalized phrases** like *fixed expressions* and *semi-fixed expressions* consist of word combinations whose overall meanings are typically not transparent from their components, e.g. *to rack one's brains* over and *to kick the bucket*. They are thus both opaque and non-compositional monolingually, but in some cases, like *elefante blanco* 'white elephant', they may be bilingually compositional.

Proverbs and similar sentential or semi-sentential constructions like adages, maxims and dicta express a general truth, belief or a moral. They are often idiomatic, opaque and non-compositional, e.g. *donde el Diablo perdió la camiseta* 'the ends of the earth'.

Collocations are recurrent combinations of words co-occurring more often than by chance whose meaning are (mostly) compositional and transparent, e.g. *mal informado* 'misinformed, incorrectly informed' (**see 5. Collocations** below)

Locutions in the context of MWEs are grammatical collocations whose central component is a function word or adverb. An example of an adverbial locution is *lo antes posible* 'as soon as possible'. Locutions are often idiomatic and non-compositional.

Multiword proper nouns, designate a person, place, company, organization, book titles and the like, e.g. *New York, George Washington, United Nations* (**see 6. Multiword proper nouns** below)

Noncontiguous MWEs are discontinuous lexical constructions that consist of fixed sequences of words interrupted by one or several gaps filled in by interchangeable words, *the more...the more*. Under this category can also be included MWEs like *be in control of*, which can be interrupted by lexical insertion, as in *be in* complete *control of*, or verbal phrases like *take off* in *he took* his jacket *off*. Non-contiguous MWEs are more challenging to identify and interpret than ordinary MWEs.

2.4 Inclusion criteria

Ideally, every type of MWE, especially non-compositional ones, should be included as headwords in both dictionaries for humans and in MT lexicons. Traditionally, dictionaries and MT lexicons have poor coverage of such MWEs as proverbs, locutions, and idiomatic expressions. It is self evident that if non-compositional MWEs are not included, or are not identified and interpreted correctly by some other means, translation accuracy will suffer.

3 Free Word Combination

3.1 Definition

A **free word combination** (FWC) is a *meaningful* free sequence of words that follow the rules of syntax but has no lexical status. FWCs have three characteristics: (1) they are potentially infinite in number, (2) they can be generated by native speakers spontaneously, and (3) they have no lexicographic or lexical status. Some examples include:

drink water
cerrar con las manos
cabrir un agujero
abrir la luz
write a poem
don't come home

FWCs are *meaningful* combinations of words (free word syntagmata), whereas meaningless combinations such as "went to New" as part of "went to New York" are ignored in linguistic analysis. FWCs are not lexical units in their own right but often appear in dictionaries in describing culture-bound terms and *untranslatable* words in place of a translational equivalent.

3.2 Analysis

It is important to note that such combinations as:

Table 3. Commbinations. n.1.

Abrir un agujero	dig a hole
Abrir un túnel	dig a tunnel
Abrir la luz	turn on the light
Abrir el agua	turn on the water

may look like MWEs, perhaps because they are not based on the primary sense of *abrir* 'to cause to open'. Nevertheless, they are indeed FWCs and have no lexical status, no more than combinations based on the primary senses of *abrir*, such as:

Table 4. Combinations n.2.

Abrir la puerta	open the door
Abrir un hospital	open a hospital
Abrir el baile	begin the dance

Such FWCs are 100 percent transparent (compositional) and productive because *abrir* is a polysemous lexeme that has such senses as 'cause to open', 'begin' and 'switch on'. The examples given are merely instances of how *abrir* is combined with direct objects.

It is important to understand this issue on the basis of objective linguistic factors, rather than subjective intuition, which is sometimes used by lexicographers in selecting dictionary entries or subentries. Below is a linguistic analysis that demonstrates that *abrir la luz* and *abrir un agujero* are indeed FWCs, rather than MWEs.

Analyzing the semantic components of *abrir* 'switch on' and *abrir* 'dig' in relation to the free word syntagmata *abrir la luz* and *abrir un agujero*, what is required to explain the need for la luz and un agujero is not, as some lexicographers may be tempted to do, to consider them integral parts of lexicalized

compound verbs, but to consider them to be semantic components consisting of an obligatory complementation of the verb by a noun phrase direct object with the selectional restriction that the complements are members of the semantic subdomain of utilities (gas, water, light...).

The fact that the senses in questions, i.e. 'dig' and 'switch on', are not central to the lexeme *abrir* is irrelevant. That is, *abrir* is a polysemic lexeme, and such productive senses as 'switch on' behave syntactically and grammatically exactly like the core meaning 'cause to open' in *abrir la puerta* 'open the door.' In other words, one must not be mislead by the peripherality of the sense 'switch on', which may make *abrir la luz* look like a collocation or compound word, rather than the FWC that it actually is.

3.3 Inclusion criteria

Such frequently co-occurring syntactic constructions like *abrir la puerta* and *abrir el baile* must not be indiscriminately considered as MWEs, though they seem to behave like lexical units. They should *not* be included in dictionaries for humans, except as example sentences, or as part of occasional "descriptive equivalents" for difficult-to-translate headwords. If such FWCs were included, dictionaries would grow to astronomical proportions since it would allow billions of meaningful FWCs.

Let us take *abrir la puerta* as an example. Since the number of potential direct objects (*ventana, entrada, boca...*) is open ended (could be extremely large), it obviously makes no sense to list them exhaustively, especially not in dictionaries for humans. Any systematic attempt to do so would bloat the dictionary out of all proportion, since the potential number of FWC can be extremely large. That is, statistically significant co-occurrences of words combinations like FWCs are syntactic constructions that do not qualify as lexical units, not only because they are completely compositional, but also because they are often highly semantically productive. On the other hand, such FWCs could serve as useful example sentences in human dictionaries.

Though compositional FWCs, which are potentially infinite in number, need not (in fact cannot) be listed in dictionaries for humans, there is one exception. If a FWC has both non-compositional and compositional translation equivalents, for the sake of clarity both compositional and non-compositional) should be included. For example, *estar por* has the compositional (literal) equivalent 'to be for' (*estar* 'to be' + *por* 'for') and the non-compositional idiomatic sense of 'to be on the verge of'.

Although FWCs such as *abrir la luz* and *abrir la puerta* are unnecessary in dictionaries for humans, they nevertheless can play a useful role in MT lexicons. Theoretically, MT systems can correctly translate such FWCs as *abrir la luz* by word sense disambiguation (WSD) even if they are not in the lexicon. That is, once the system determines that the sense of *abrir* in this context is 'switch on', it can correctly translate to 'turn on the light'. Nevertheless, since memory is virtually unlimited, it makes sense to include some high-frequency FWCs in the MT lexicon explicitly because it greatly simplifies processing; that is, a simple

lookup operation replaces the sophisticated semantic and contextual analysis that is required for WSD.

4 Phraset

4.1 Definition

A phraset is a free, *meaningful* combination of words (FWC) that is recurrently used to express a concept that has no lexical status but corresponds to a lexical unit in another language, e.g. *cerrar con llave* correponds to 'lock' and *ir en bicicleta* correponds to 'cycle'.

Bentivogli and Pianta[1] [2] have discussed phrasets in detail in the context of WordNet. The term as used here has the following characteristics: (a) syntactically and grammatically they are indistinguishable from FWCs, (b) they have no lexical status but correspond to lexical units in another language, and (c) they are often used in bilingual dictionaries as a "descriptive equivalent" for lexical gaps, e.g. *cerrar con llave* as the equivalent of the lexical unit 'lock'.

4.2 Analysis

Lexical *anisomorphism*, a basic feature of language, refers to the lexical incompatibility between languages. One manifestation of this is the large number of *lexical gaps* in every language; that is, words that have no equivalents in the target language. Bilingual dictionaries overcome this by providing descriptive equivalents when possible, similar to a definition in monolingual dictionaries.

It is important to note that phrasets are not "lexical units" in the source language and are not normally listed as headwords in dictionaries (though they may appear as subentries or in example sentences) since their status in the language is essentially the same as FWCs. However, many phrasets do behave like lexical units; that is, they have semantic integrity and cohesiveness and express a concept compositionally for which the language lacks an established lexical unit.

It is only when viewed from the point of view of the *target* language that phrasets acquire a special status. For example, English-Spanish dictionaries translate *to cycle* by the phraset *ir en bicicleta*, so it gets its special status, if we can call it that, by virtue of that fact alone, not because native speakers consider it "special" in any way. This demonstrates an interesting and useful fact about phrasets: that they can be monolingually compositional and transparent (as *ir en bicicleta*), yet bilingually non-compositional or a simplex lexical units (ascycle).

Distinguishing between FWCs and phrasets is, in principle, very difficult and often impossible since monolingually they behave identically. For example,though *write a poem* is an FWC in English, in Japanese there is a verb 作詩する *sakushi suru*, translated as *'write a poem'* in English, so that write a poem can be classified as a phraset, rather than an FWC, from a Japanese point of

view. For native speakers of English, write a poem has no special status – that is, it has exactly the same status as *write a letter, write a song, write a book* etc. – and consequently will not appear as a dictionary entry even in the most comprehensive monolingual English dictionaries, nor as a source headword in English-to-X bilingual dictionaries.

Why is this so? For native speakers, phrasets do not have a psychological reality as a combination of words that need to be treated as meaningful units; that is, they are completely transparent and compositional and thus are (probably) not registered in the internal mental lexicon of native speakers.

It should be noted that phrasets, just like FWCs and phrasal verbs (which are full-fledged lexical items), can be noncontiguous: that is, the phraset *cerrar con llave* can be interrupted by lexical insertion, as in *cerrar la puerta con llave*, adding to the difficulty of detecting them.

Phrasets are very useful for translating lexical gaps into the target language. Statistical techniques, such as extracting contiguous bigrams and trigrams of high occurrence and high semantic cohesiveness, have been used to detect them, but because monolingually their behavior is identical to that of FWCs, this is a difficult task. For example, *cerrar con llave*is identical in structure to *cerrar con las manos* – that is, they both have identical surface structures. One effective technique for detecting phrasets is to reverse the entries of bilingual dictionaries. For example, the entry lock in an English-Spanish dictionary will yield the phraset *cerrar con llave*. Another technique is using a database of bilingual aligned example sentences found in bilingual dictionaries, which are an excellent source of phrasets, or using sentence aligned parallel corpora.

4.3 Inclusion criteria

Most dictionaries for humans rarely, if ever, intentionally include phrasets as source language entries or subentries, since phrasets are semantically compositional and have no lexical status. On the other hand, MT lexicons, designed to achieve full reversibility and comprehensive coverage, listing phrasets is not only desirable but essential for achieving high translation accuracy. There is no question that including Spanish phrasets like *cerrar con llave* 'to lock' in a Spanish-English MT lexicon is essential since these cannot be compositionally translated into English (the literal translation 'close with a key' is unidiomatic and incorrect).

Another good example of a phraset is *ir en bicicleta*, which is equivalent to the English lexeme 'to cycle' and the FWC 'ride a bicycle'. *Ir en bicicleta* is not normally listed as a source language entry in Spanish dictionaries; if it were, the question would be *where to draw the line*? That is, why not also include:

ir en coche
ir en carro
ir en monociclo
ir en avión
ir en patines

ir en globo

All of these have exactly the same status in Spanish, linguistically, lexicographically, and psychologically in the minds of native speakers and are thus are no different from *ir en bicicleta*.

It is worthwhile noting that translating source-language lexical units non-compositionally into target language phrasets is common, especially between highly anisopmorphic language pairs like Japanese and English. This problem cannot be solved by haphazardly listing phrasets in dictionaries, but requires a comprehensive approach in which phrasets are collected systematically.

It is further worthwhile noting that grammatical anisomorphism combined with lexical anisomorphism are the reasons why languages have many "untranslatable" lexical units, some of which are not just "difficult to translate" but *in principle completely impossible to translate*. For example, *would've* as an isolated word is impossible to translate into many languages. Dictionaries *describe*, rather than translate, such words by using FWCs or phrasets. Because of the lexical, conceptual, and grammatical differences between languages, some lexical units in a language cannot, *in principle*, be translated, not because of lack of lexicographic skill.

5 Collocation

5.1 Definition

A **collocation** (or *institutionalized phrase*) is a recurrent combination of words that co-occur more often than by chance whose meaning is (mostly) compositional and transparent.

Table 5. Collocations

Bonita sorpresa	nice surprise
Estar fascinado con	be fascinated with
Tomar una decisión	make a decision
Hacer una pausa	take a break
Prestar atención	pay attention
Hacer amor a/con	make love to/with
Respecto a	With regard to
Abandonarse a la desperación	to fall into despair

Collocations are a subtype of MWEs that have the following criteria: (a) they often cannot be translated literally, (b) they are (mostly) compositional and semantically transparent, (c) their components cannot be replaced without losing idiomaticity, and (d) they do not have full lexical status (see Table 5).

5.2 Analysis

A collocation is a group of words that co-occur more frequently than by chance. Collocations are a phenomenon that have linguistic status, and are also useful for statistical analysis and natural language processing. However, with the exception of specialized dictionaries of collocations and idiomatic phrases, they do not normally appear in dictionaries as main entries or subentries, but may appear in example sentences.

Collocations are difficult to define precisely. They are often discussed in contrast with FWCs on one end of the spectrum and with idiomatic expressions on the other. Whereas FWCs can be described in terms of general syntactic rules and semantic restrictions, idioms are fixed word combinations that are difficult or impossible to generalize. Collocations fall between these two extremes, though drawing a clear line between them is not always possible. Studying the above examples carefully should clarify how collocations differ from full-fledged lexical units.

Collocations don't have full lexical status because they don't normally represent concepts. For example, *nice surprise* is a conventional phrase corresponding to *bonita sorpresa* (*beautiful surprise* would be unidiomatic), but cannot be considered to be a full-fledged lexical unit in either language. That is, one can say that they are probably not registered in the brain's internal lexicon. They are more in the realm of usage conventions rather than full-fledged lexical units.

5.3 Inclusion criteria

Regardless of the theoretical distinction between collocations and full-fledged lexical units, because collocations are so common it is desirable to include them in bilingual dictionaries, not to speak of MT lexicons, in order to achieve higher translation quality.

Since collocations are compositional and semantically transparent, there is some chance that a human or MT system can translate them correctly by word-for-word substitution, but it is nevertheless desirable to list them explicitly in order to achieve better, unambiguous result.

It is highly desirable to make a systematic effort to collect collocations for including in comprehensive bilingual dictionaries as well as in MT lexicons. One technique for acquiring collocations is to extract them from aligned example databases based on paper dictionaries; another is to extract them from bilingual parallel corpora.

Table 6. Inaccurate translations of POIs.

Japanese	Google	Bing	Baidu	NICT	CJKI
海の中道線	Midair line of the sea	The middle line of the sea	The sea line	海の中道線	Umi-no-Nakamichi Line
三角線	Triangle	Triangular line	Misumi	Misumi Line	Misumi Line
十和田観光電鉄線	Triangle	Triangular line	Misumi	Misumi Line	Misumi Line
神津島空港	Towada Shimbun photoelectric wire	Towada Kanko railway line	Towada sightseeing electric railway line	Towada Kankō Electric Railway Line	Towada Kanko Electric Railway Line
神津島空港	Kozu Island airport	God Tsushima Airport	Kozu Island Airport	Kōzushima Airport	Kozushima Airport
中部国際空港	Chubu International Airport	Chubu International Airport	Central Japan International Airport	Chubu International Airport	Chubu Centrair International Airport
鬼の城公園	Demon Castle Park	Demon Castle Park	Demon Castle Park	Oni Castle Park	Oninojo Park

6 Multiword Proper Nouns

6.1 Definition

A *multiword proper noun* is a combination of two or more words that together function as a single proper noun. This includes place names such as *Republic of China*, companies and organizations such as *United Nations*, personal names such as *Shinzo Abe*, and points of interest (POI) such as *Narita International Airport*.

6.2 Analysis

The recognition and accurate translation of proper nouns, many of which are bilingually non-compositional, are a major issue in MT and other NLP applications. This is especially true for Chinese and Japanese, whose scripts present linguistic and algorithmic challenges not found in other languages. These difficulties are exacerbated by the lack of easily-available comprehensive lexical resources for proper nouns, especially POIs, resulting in a high rate of translation failure.

The CJK Dictionary Institute (CJKI), which specializes in CJK and Arabic computational lexicography, has been engaged in the construction of large-scale lexical resources that cover tens of millions personal names, place names, and POIs. These resources and methodology are described in Halpern[6] and

Halpern[5]. Tests have shown that MT systems, including those using neural networks, often fail to accurately translate proper nouns, especially POIs. For example, a test to translate 75 Japanese POIs gave surprisingly poor results, as shown in Table 1. For example, 鬼の城公園 is translated by Baidu and Google word for word as 'Demon Castle Park', whereas the actual name of this park is 'Oninojo Park'.

6.3 Inclusion criteria

Dictionaries for humans normally do not include proper nouns, except possibly for very well know place names such as country names and famous people. Human users do not expect, and have no need for, comprehensive coverage of proper nouns. MT lexicons, on the other hand, should include as many proper nouns as possible. In fact, most MT systems perform poorly in translating proper nouns in general and multiword POIs in particular. To achieve higher translation accuracy, proper noun resources for MT must be greatly expanded.

7 Conclusions

This paper attempts to define the various types of MWUs, and to clarify the underlying linguistic concepts on the basis of linguistic and lexicographic principles while taking into consideration the needs of MT lexicons. It is clear that the accurate identification and translation of MWUs are critical to enhancing the translation accuracy of MT systems. It is hoped that the analysis given here will contribute to the improved identification of MWUs, based on (mostly) objective criteria, and that MT system developers will pay greater attention to the importance of large-scale lexicons with comprehensive coverage of MWUs, including proper nouns.

References

1. Bentivogli, L., Pianta, E.: Beyond lexical units: enriching wordnets with phrasets. In: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 2 (EACL)., vol. 2, pp. 67-70. Association for Computational Linguistics, Stroudsburg, PA, USA (2003). DOI: https://doi.org/10.3115/1067737.1067750
2. Bentivogli, L., Pianta, E.: Extending WordNet with Syntagmatic Information. In: Sojka,P., Pala, K., Smrz, P., Fellbaum, C., Vossen, P. (eds.), Proceedings of the Second International WordNet Conference - GWC 2004, pp. 47.53. Masaryk University Brno, Czech Republic, (2004).
3. Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., Zampolli, A.: Towards best practice for multiword expressions in computational lexicons. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), pp. 1934-1940. (2002) http://www.lrec-conf.org/proceedings/lrec2002/pdf/259.pdf.

4. Carbonell, J., Klein, S., Miller, D., Steinbaum, M., Grassiany, T., Frey, J.: Context-based machine translation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas. (2006) http://www.mt-archive.info/AMTA-2006-Carbonell.pdf.
5. Halpern, J.: The Role of Lexical Resources in CJK Natural Language Processing. In: Proceedings of the Fifth Workshop on Chinese Language Processing, SIGHAN@COLING/ACL 2006, pp. 22-23. Association for Computational Linguistics 2006, Sydney, Australia (2006).
6. Halpern, J.: Very Large-scale Lexical Resources to Enhance Chinese and Japanese Machine Translation. In: TAUS Executive Forum Tokyo 2017. (2017).
7. Henderson, J. A.: What's in a Word?. The University of Edinburgh, Edinburgh (2007).
8. Sag, I. A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh, A. F. (Ed.) Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), pp. 1-15. Springer-Verlag, Berlin, Heidelberg (2001).
9. Váradi, T.: Multiword units in an MT lexicon. In: Proceedings of the Workshop on Multi-word-expressions in a multilingual context. (2006) https://www.aclweb.org/anthology/W06-24.
10. Villavicencio, A., Bond, F., Korhonen, A., McCarthy, D.: Editorial: Introduction to the special issue on multiword expressions: Having a crack at a hard nut. in: Comput. Speech Lang. 19, 4 (October 2005), pp. 365-377. (2005) http://dx.doi.org/10.1016/j.csl.2005.05.001.