The Framework of Multiword Expression in Indonesian Language

Totok Suhardijanto
Department of Linguistics
Faculty of Humanities
Universitas Indonesia
totok.suhardijanto@ui.ac.id

Rahmad Mahendra
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
rahmad.mahendra@cs.ui.ac.id

Zahroh Nuriah
Department of Linguistics
Faculty of Humanities
Universitas Indonesia
zahroh.nuriah@ui.ac.id

Adi Budiwiyanto
Department of Linguistics
Faculty of Humanities
Universitas Indonesia
adi.budiwiyanto71@ui.ac.id

Abstract

This paper presents our attempt to develop an Indonesian multi-word expression (MWE) identification framework. The framework consists of three different steps. In the first step, we surveyed any definitions and categorizations of MWEs in the previous studies. In the second step, sentences from our language corpora are segmented and high-frequency n-grams are extracted using statistical methods. The extraction results which consist of word sequences are evaluated and reorganized by using the phraseological analysis procedure. This procedure consists of polylexicality, fixedness, and idiomaticity. The final step is to recategorize and reevaluate data collected in the second step. In this research, data is collected from newspapers, Wikipedia, and scientific papers. The result shows that in terms of Indonesian MWEs, despite polylexicality and fixedness, idiomaticity should be correlated with compositionality to get a better classification of MWEs.

1 Introduction

Since Firth (1957) stated that in language communication, meaning is usually conveyed by word group, many scholars had explored this idea and had ended up with a concept of regularly significant frequent sequences of words called multiword expressions. Different scholars used different terms for these sequential linguistic phenomena (see Burger, 2015; Fleischer, 1997; Sprenger, 2003; Biber & Conrad, 2019). Burger (2015) and Fleischer (1997) mentioned it as phraseology, while Sprenger (2003) and Bider & Conrad (2019), called these word sequences as fixed expressions. Burger (2015) distinguishes phraseological units or MWEs by three features: polylexicality, fixedness, and idiomaticity, where idiomaticity need not be present in all phraseological units.

Up to now, in the field of linguistics and language studies, the classification of MWEs is still considered as a challenging task. Even more so, as there is no general consensus about what counts as an MWE. Masini (2005: 145) viewed MWEs as “lexical units larger than a word that can bear both idiomatic and compositional meanings. (···) the term multi-word expression is used as a pre-theoretical label to include the range of phenomena that goes from collocations to fixed expressions” . In a more detailed way, Sailer & Markantonatou (2018: v) defines MWEs as ‘any expression that contains more than one basic lexical element and that is lexicalised, fixed, idiomatic, or irregular in one way or the other.’

For both the natural language application and the linguistic theory, Multiword expressions (MWEs) are challenging because they are often difficult to be classified by an application of the machinery developed for free combinations where the default is that the meaning of an utterance can be predicted from its structure. Compared to MWEs in European languages, there are only a few number of primarily descriptive works on MWEs for Asian languages, except for Chinese, Japanese and Korean. This paper contributes to look at MWEs in Indonesian. They discuss prominent issues in MWE research such as classification of MWEs, their formal grammatical modeling, and the description of individual MWE types from the point of view of different theoretical frameworks, such as Dependency Grammar, Generative Grammar, Head-driven Phrase Structure Grammar, Lexical Functional Grammar, Lexicon Grammar.

Why do we need to identify and deal with MWEs in natural language processing? There are several reasons that can be explained further. First, it is important to recognize MWEs first before processing and implementing POS tagging. Secondly, in current NLP researches, when a cross-lingual or multilingual approach is adopted to a machine learning project, it needs to identify whether the counter translation in a language is a simple or a complex word? Furthermore, identifying opaque or more idiomatic MWEs such as kick the bucket is also a challenging task, but the task can be very helpful when it comes to improving automatic sentiment analysis of data collected from social networks.

In this paper, our research objectives consist of twofold. First, this study attempted to identify MWEs in Indonesian. Second, this paper also classifies Indonesian MWEs into categories with the same distribution and behaviors. [introduction stub]

2 Classification of MWEs

Although some of the classifications made by computational linguists appear to be different, there are actually similarities in the categories of the classifications. The classifications they developed rely on the previous works. Sag et al. (2002) classify MWEs into (1) **institutionalised phrases**, i.e., sets of words which co-occur often but have no syntactic idiosyncrasy, and whose semantics are fairly compositional, and (2) **lexicalised phrases**, presenting some idiosyncratic syntactic or semantic characteristics. The latter can be further divided into three subclasses according to their degree of flexibility: fixed (e.g., *ad hoc, vice versa*), semi-fixed and syntactically flexible expressions (e.g., *cable car*), and proper names (*San Francisco*), as well as non-decomposable idioms (e.g., *kick the bucket, shoot the breeze*).

Meanwhile, Baldwin et al. (2010) categorised MWEs into three types: (1) **nominal MWEs**, (2) **verbal MWEs** and (3) **prepositional MWEs**. Nominal MWEs are one of the most common types (e.g., *golf club*, *connecting flight*, or *open secret*). The verbal MWE consists of (a) verb-particle construction, also termed *particle verbs* or *phrasal verb*, (e.g., *play around*, *take off*, *cut short* and *let go*); (b) prepositional verbs, (e.g., *refer to*, *look for*, *come across*, and *grow on*); (c) Light verb constructions (e.g., *do a demo*, *give a kiss*, *have a drink*, and *take a walk*; and (d) Verb-noun idiomatic combinations (VNICs, also known VP idioms), e.g. *kick the bucket*, *shoot the breeze*, and *spill the beans*. The prepositional MWEs comprises (a) determinerless-prepositional phrases (e.g. *on top*, *by car*, and *high expense*) and (b) complex prepositions (e.g. *on top of*, *in addition to*, and *with regard to*).

Based particularly on the syntactic and semantic properties, MWEs can be classified into (1) **lexicalised phrases** and (2) **institutionalised phrases**. Lexicalised phrases are MWEs with lexical, syntactic, semantic or pragmatic idiomaticity. Lexicalised phrases can be further split into a) fixed expressions (e.g., *ad hoc*, *at first*), *semi-fixed expressions* (e.g., *spill the beans*, *car dealer*, *Chicago White Socks*) and syntactically-flexible expressions (e.g. *add up*, *give a demo*). On the other hand, the class of institutionalised phrases corresponds to MWEs which are exclusively statistically idiomatic (e.g., *salt and pepper*, *many thanks*).

Ramisch (2015: 42-44) makes a simplified typology based on (1) the morphosyntactic role of the whole expression in a sentence and (2) its difficulty to be dealt with using computational methods. They divide MWEs into three categories: (1**) nominal expressions** consisting of nominal compounds (e.g., *traffic light, Russian roulette, degree of freedom, wine glass*), proper names (e.g., *United Nation, Porto Alegre, Alan Turing*), and multiword terms; (2) **verbal expressions** comprising (a) phrasal verbs which are divided into transitive prepositional verbs (e.g., *rely on, agree with*) and more opaque verb-particle constructions where the particle is actually attached to the verb, forming a cohesive lexical-semantic unit (e.g., *give up, take off*), and (b) light verb constructions (e.g., *take a shower, make a presentation*); and (3) **adverbial and adjectival expressions** (e.g., *upside down; second hand, on fire, at stake,* and *in the buff*).

In addition to those three main types, they also define three orthogonal types that are more related to the computational methods used to treat MWEs. The first class is **fixed expressions** that can be dealt with using relatively simple techniques. They correspond to the fixed expressions of Sag et al. (2002). The second class is **idioms**, that are very hard to recognise and require the use of external semantic resources. The last class is called **true collocations** that correspond to the notion of words that co-occur more often than expected by chance.

Müller et al. (2015: 669) give a list of MWEs though it is not intended as a classification of MWE: (1) proverbs (A bird in the hand is worth two in the bush), quotations (Shaken, not stirred) and commonplaces (One never knows); (2) metaphorical expressions (as sure as eggs is eggs); (3) verbal idioms (to kick the bucket); (4) particle/phrasal verbs (to make up); (5) light verb constructions/composite predicates (to have a look); (6) syntactic/quasi noun incorporation (to wash car, to play piano, to buy (a) house, to have/own a car); (7) stereotyped comparisons/similes (as nice as pie, swear like a trooper); (8) binomial expressions (shoulder to shoulder, by and by, nourish and cherish); (9) complex nominals (man about town, weapons of mass destruction, sheep's clothing) - Collocations (strong tea, hard frost); (10) fossilized/frozen forms (all of a sudden, instead of, depending on); (11) routine formulas (Good morning, How are you doing?, Happy Birthday). Particularly in French language, Müller et al. (2015) classify MWEs into seven categories: (1) nominal sequence, (2) verbal sequence, (3) adjectival sequence, (4) adverbial sequence, (5) prepositional and conjunctive sequence, (6) determinative sequence, and (7) multi-word interjection.

Constant et al. (2017: 6-8) make a list of MWE categories commonly seen in the literature. The categories are non-exhaustive and can overlap. They cover **idioms** (e.g., *to kick the bucket*), **light-verb constructions** (e.g., *to take a shower*), **verb-particle constructions** or phrasal verbs (e.g., *to give up*), **compounds** (e.g., *dry run, bank robbery, stir fry*), **complex function words** (e.g., *as soon as, up until, by and large*), multiword named entities (e.g., *International Business Machines*), and multiword terms (e.g., *short-term scientific mission*).

Laporte (2018) adapts Baldwin and Kim (2010) classification. He proposes two types of classification which are different in terms of the existence of copula in the support-verb constructions. His classification is based on comparison of MWEs in some languages, such as English, French, Romance, Greek, Arabic, Chinese or Korean.

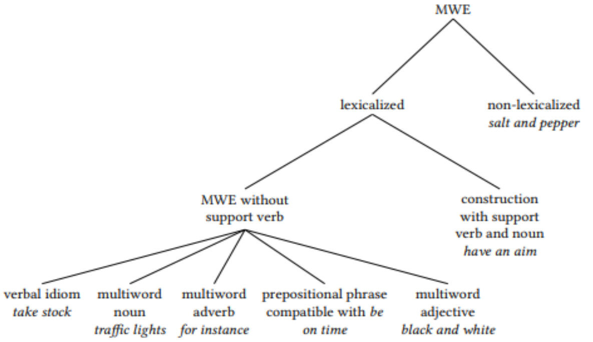

Figure 1: Classification by Laporte (2018) (Escartín et al., 2018)

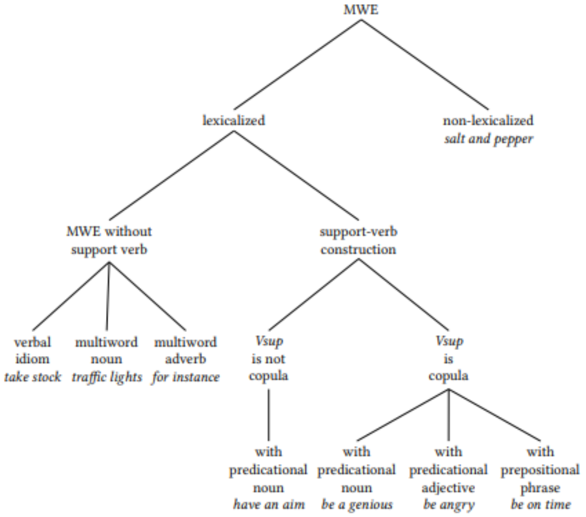

Figure 2: Classification by Laporte (2018) where copula is considered a support-verb (Escartín et al., 2018)

3 MWE Framework in Indonesian

Fleischer (1997) and Burger (2015) mentioned the same three prototypical properties of MWEs: polylexicality, fixedness, and idiomaticity. However, Fleischer proposed one more prototypical property that is not implemented in this study: lexicalization. In our framework, since lexicalization is quite rare in Indonesian MWEs, we did not take it into account to identify and classify MWEs in this paper. According to Fleischer, as MWE is a fuzzy concept, polylexicality is the most obligatory property, while Burger wrote that idiomaticity need not be present in all types of MWEs.

In our framework, MWEs are evaluated from three properties: polylexicality, fixedness, and idiomaticity. First, all MWEs should be made up of more than one word such as *Susilo Bambang Yudhoyono*, *singa laut* 'walrus', and *rumah sakit* 'hospital'. Second, most MWEs cannot be modified such as (i) be inserted by one or more words: *rumah sakit* 'hospital' cannot be inserted with yang 'that' (relative pronoun) become *rumah yang sakit* which literally means 'a sick house'; (ii) be reversed in terms of word order: *singa laut* 'walrus' cannot be reversed into *laut singa* which literally means 'the Lion Sea'. With regard to idiomaticity, we agree with Fleischer and Burger that it is not obligatory to determine whether a word sequence is a MWE or

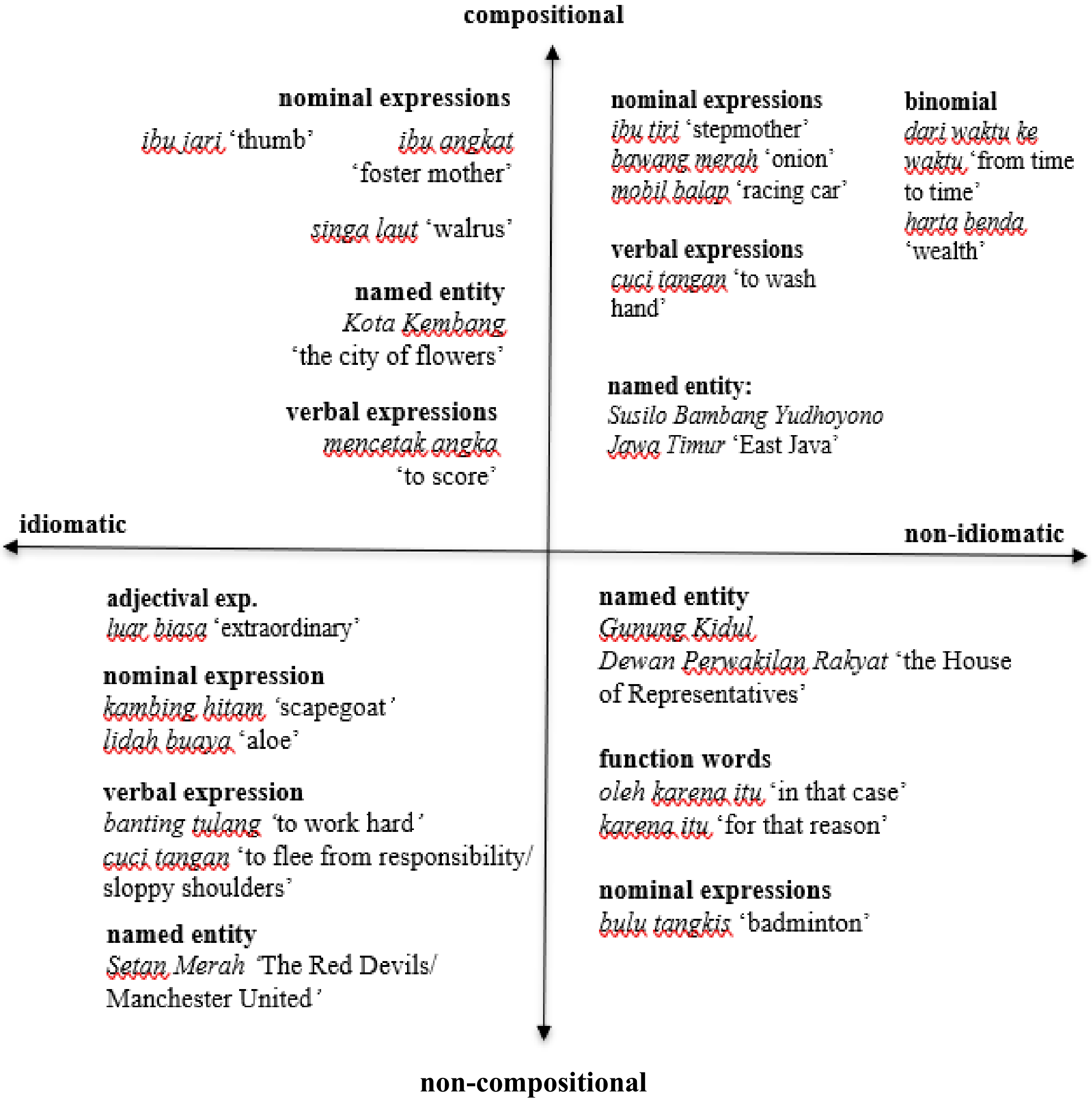

Figure 3: MWE classification based on idiomaticity and compositionality

not. However, in dealing with Indonesian MWEs, we need to consider the idiomaticity as a complex property, rather than a single property. We need to consider the structure of components among MWEs to deal with idiomaticity in Indonesian MWEs.

According to Cruys and Moiron (2007), the linguistic behaviour of MWEs can not be predicted based on the linguistic behaviour of their component words. In addition, Baldwin (2006) characterizes the idiosyncratic behavior of MWEs as “a lack of compositionality manifested at different levels of analysis, namely, lexical, morphological, syntactic, semantic, pragmatic and statistica.” MWEs in Indonesian also show idiosyncratic behavior, but some of those still manifest compositionality. Therefore, we classify MWEs in Indonesia based on two axes, that is, idiomaticity and compositionality.

If we look closer at the components, the characteristics of MWEs in Indonesian is not really different from those of MWEs in other languages discussed in the former studies.

- Verbal expressions are verbs that consist of more than one word but refer to one concept. These expressions can consist of a noun, adjective, preposition or other verb next to another verb. Verbal expressions can be an incorporation with a noun as the patient such as *minta izin* ‘to ask for permission’, *mencetak angka* ‘to score’, *cuci tangan* ‘to wash hand’ or ‘to flee from responsibility/sloppy shoulder’, *banting tulang* ‘to work hard’ (VN), or a serial verb like *siap saji* ‘fast food’ (VV), or a verbal combination with an adjective like *bekerja keras* ‘to work hard’ (VA), or a preposition like *tinggal di* ‘to live in’ (VP).
- Nominal expressions are nouns that consist of more than one word but refer to one concept. These expressions can consist of an adjective, verb or other noun next to another noun. For example *bawang merah* ‘onion’, *kambing hitam* ‘scapegoat’, *ibu tiri* ‘stepmother’ (NA), *meja makan* ‘dining table’, *ibu angkat* ‘foster mother’, *mobil balap* ‘race car’, *bulu tangkis* ‘badminton’ (NV), *kata kunci* ‘keyword’. The second nouns of nominal MWEs that consist of nouns (N1 N2) semantically can have agentive (*belanja negara* ‘state expenditure budget’), patientive (*uji korelasi* ‘corellation test’), genitive (*lidah buaya* ‘aloe’, *ibu jari* ‘thumb’) or benefactive (*ruang publik* ‘public area’) or other relation with the first noun (*singa laut* ‘walrus’). Some nominal expressions are named entities that consist of a noun and other words (N X(X)) such as *Kota Kembang* ‘the City of Flower’, *Jawa Timur* ‘East Java’, *Setan Merah* ‘Red Devils/Manchester United’, *Dewan Perwakilan Rakyat* ‘The House of Representatives’, *Gunung Kidul* ‘South Mountain’, *Susilo Bambang Yudhoyono* (N(X)). There are also binomial with two words or more like *harta benda* ‘wealth’, *waktu ke waktu* ‘from time to time’ ((P)N(P)N) and nominal expressions with three words or more *tahun kerja efektif* ‘work year’ (NVA), *defisit transaksi berjalan* ‘current account deficit’ (NNV), *berat badan lahir rendah* ‘low birth weight’ (NNVA).
- Adjectival expressions are adjectives that consist of more than one word but refer to one concept. These expressions can consist of an adjective with another adjective, for example *tegak lurus* ‘perpendicular’ (AA) or a noun with an adjective for example *luar biasa* ‘outstanding’ (NA). The construction NA is problematic because this form has no head since the head of a phrase in Indonesian is the word at the beginning (the left one).
- Adverbial expressions are adverbs that consist of more than one word but refer to one concept. These expressions can consist of an adverb with another adverb, for example *lebih kurang* ‘approximately’ (Adv Adv) or an adverb with a verb for example *lebih lanjut* ‘furthermore’ (Adv V)
- Function words that consist of more than one word are preposition *di luar* ‘outside’, *di dalam* ‘inside’ (PN) or conjunction *akan tetapi* ‘but’, oleh karena (itu) ‘therefore’ (PConj).

4 Why Indonesian NLP should care about MWEs

In dealing with identification and classification of MWEs in a particular language, specific problems that are related to the language can arise. With regard to Indonesian MWEs, problems are not only caused by the possible idiomatic meaning of the MWE, but also by the ambiguity of the POS of the MWE’s components. For instance, the MWE *minta izin* ‘ask permission’ consists of the word *izin* ‘permission’ that is a noun in that case, but in nonformal context, such as in spoken Indonesian, *izin* can also appear as a verb (*Saya izin datang terlambat* ‘I was allowed to come late’*)*.

Another example is shown in the following case. The rule: NP → N V is actually not a valid context-free grammar rule because a NP sequence usually consists of two or more nouns: N N (*batu pasir* ‘sandstone’) or N N N (*batu pasir Navajo* ‘Navajo sandstone’). Meanwhile, a sequence with N followed by V is usually treated as a clause rather than a phrase, such as in *Jakarta banjir* (NV) which means ‘Jakarta floods’. However, in another NV sequence, such as *rumah makan* ‘restaurant’, it is considered as a phrase and categorized as a MWE. If MWEs are not processed first, *rumah* ‘house’ and *makan* ‘eat’ are treated as two separated words. The first word would be tagged with the noun category, while the latter would be tagged with the verb category.

Suppose that MWE is not processed, the meaning of *rumah makan* may be derived by applying some inference rules when concatenating the words *rumah* and *makan*”. The word *rumah makan* ‘restaurant’ can be defined as ‘the house where to eat’. However, this heuristic should fail in another case like “rumah sakit” (*hospital* in English). If we apply the same rule as previously, combining the word “rumah” and “sakit” (*sick* in English) may come up with a meaning of the house where to be sick.

On the other hand, in spite of following the same pattern with phrase constructions in Indonesian language grammar, several MWEs are idiomatics that have implications of totally different meaning. For example: *kambing hitam* ‘*scapegoat*’ vs *anjing hitam* ‘black dog’. So, for those reasons discussed above, the processing of MWE is an ideal step to lighten the burden of lexical semantics processing.

5 Conclusion

In this paper we proposed the framework for identification and classification of Indonesian MWEs. We evaluated MWEs in Indonesian by implementing three characteristics: polylexicality, fixedness, and idiomaticity, to categorize Indonesian MWEs. However, in order to determine a sequence as a MWE in Indonesian, we need to correlate idiomaticity and compositionality to make the classification result more clear. MWE processing should benefit the natural language processing tasks in Indonesian language. We have illustrated a few examples in which ignoring the MWE phenomenon can lead to such problems in POS Tagging, Syntactic Parsing, and Lexical Semantics Processing.

In a future work, we will study how to identify other MWE properties, such as discontinuity and variability, for Indonesian that have not been discussed in this paper. We also expect to extend our work to extract MWE lexicon from large corpora in Indonesian language and incorporate Indonesian MWE lexicon to other language resources and use them in Natural Language Processing tasks. Another interesting direction to further investigate is to include Indonesian MWEs from spoken registers because we found that MWEs in spoken Indonesian are more challenging to deal with.

References

Carla Parra Escartín, Almudena Nevado Llopis, and Eoghan Sánchez Martínez. 2018. Spanish multiword expressions: Looki*ng* for a taxonomy. In Manfred Sailer & Stella Markantonatou (eds.)*, Multiword expressions*: *Insights from a Multi-lingual Perspective*, 271–323. Berlin: Language Science Press. DOI:10.5281/zenodo.1182597.

Carlos Ramisch. 2015. *Multiword expressions acquisition*: *A generic and open framewor*k. Cham/Heidelberg/New York/Dordrecht London: Springer.

Éric Laporte. 2018. Choosing features for classifying multiword expressions. In Manfred Sailer and Stella Markantonatou (eds.), *Multiword expression*s: *Insights from a multi-lingual perspective*, 143–186. Berlin: Language Science Press. DOI:10.5281/zenodo.1182597

Harald Burger. 2015. *Phraseologie: Eine Einführung am Beispiel des Deutschen*. 5th edn. Berlin: Erich Schmidt Verlag.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics* (CICLing-2002), 1–15

Manfred Sailer and Stella Markantonatou. 2018. *Multiword Expressions: Insights from a Multi-lingual Perspective*. London: Language Science Press.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics* 43, no. 4 (2017): 837-892.

Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer. 2015. *Word Formation*: *An International Handbook of the languages of Europe.* Vol. 1. Berlin/Boston: De Gruyter Mouton.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. Cambridge, MA: MIT Press.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya & Fred J. Damerau (eds.), *Handbook of Natural Language Processing*, 2nd edn., 267–292. Boca Raton: CRC Press.

Van de Cruys, T. and Moirón, B.V., 2007, June. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions (pp. 25-32).*

Wolfgang Fleischer. 1997. *Phraseologie der deutschen Gegenwartssprache*. 2nd edn. Tübingen: Niemeyer.