Multiword Expression Processing: A Survey

Mathieu Constant*
ATILF, Université de Lorraine & CNRS

Gülşen Eryiğit**
Istanbul Technical University

Johanna Monti†
"L'Orientale" University of Naples

Lonneke van der Plas‡
University of Malta

Carlos Ramisch§
Aix Marseille University, CNRS

Michael Rosner‡
University of Malta

Amalia Todirascu‖
LiLPa, Strasbourg University

Multiword expressions (MWEs) are a class of linguistic forms spanning conventional word boundaries that are both idiosyncratic and pervasive across different languages. The structure of linguistic processing that depends on the clear distinction between words and phrases has to be re-thought to accommodate MWEs. The issue of MWE handling is crucial for NLP applications, where it raises a number of challenges. The emergence of solutions in the absence of guiding principles motivates this survey, whose aim is not only to provide a focused review

* ATILF, Université de Lorraine & CNRS, 44 avenue de la Libération, 45000, Nancy, France. E-mail: Mathieu.Constant@univ-lorraine.fr.
** ITU Department of Computer Engineering, 34469, Istanbul, Turkey. E-mail: gulsen.cebiroglu@itu.edu.tr.
† "L'Orientale" University of Naples, Palazzo Santa Maria Porta Coeli, Via Duomo, 219 80138 Naples, Italy. E-mail: jmonti@unior.it.
‡ University of Malta, Tal-Qroqq, Msida MSD2080, Malta. E-mail: {lonneke.vanderplas, mike.rosner}@um.edu.mt.
§ Aix Marseille University, CNRS, LIF, 163 av de Luminy – case 901, 13288, Marseille Cedex 9, France. E-mail: carlos.ramisch@lif.univ-mrs.fr.
‖ LiLPa, Strasbourg University, 22, rue René Descartes, BP 80010, 67084, Strasbourg Cedex, France. E-mail: todiras@unistra.fr.

Submission received: 26 May 2016; revised version received: 24 March 2017; accepted for publication: 2 June 2017.

doi:10.1162/COLI_a_00302

of MWE processing, but also to clarify the nature of interactions between MWE processing and downstream applications. We propose a conceptual framework within which challenges and research contributions can be positioned. It offers a shared understanding of what is meant by "MWE processing," distinguishing the subtasks of MWE discovery and identification. It also elucidates the interactions between MWE processing and two use cases: Parsing and machine translation. Many of the approaches in the literature can be differentiated according to how MWE processing is timed with respect to underlying use cases. We discuss how such orchestration choices affect the scope of MWE-aware systems. For each of the two MWE processing subtasks and for each of the two use cases, we conclude on open issues and research perspectives.

1. Introduction

Traditional formal descriptions of individual human languages typically divide labor between a repository of words and their properties, called a **lexicon**, and a description of how such words combine to form larger units, called a **grammar**.[1] These two elements provide a systematic but finite basis for computing the properties of any syntactically legitimate sentence. Although grammatical theories differ about the nature of lexical versus grammatical information and their manner of interaction, a particular theory must establish what counts as a word in order to pin down what the lexicon should contain.

As Baldwin and Kim (2010), among others, have pointed out, the question of what constitutes a word is surprisingly complex, and one reason for this is the predominance in everyday language of elements known as **multiword expressions** (MWEs). MWEs consist of several words (in the conventionally understood sense) but behave as single words to some extent. This is well illustrated by an expression like *by and large*, which any English speaker knows can have roughly equivalent meaning and syntactic function to *mostly*, an adverb. Among the problematic characteristics of this expression are (1) syntactic anomaly of the part-of-speech (POS) sequence preposition + conjunction + adjective, (2) non-compositionality: semantics of the whole that is unrelated to the individual pieces, (3) non-substitutability of synonym words (e.g., *by and big*), and (4) ambiguity between MWE and non-MWE readings of a substring *by and large* (e.g., *by and large we agree* versus *he walked by and large tractors passed him*).

Although these characteristics by no means exhaust the list of peculiarities, the idiosyncratic nature of the expression is plain, leading us to ask where its pertinent characteristics should be stored. The traditional division of labor gives us two options—the lexicon or the grammar—but MWEs disrupt the tradition precisely because they are more than one word long (Sag et al. 2002). Their idiosyncrasy suggests that they belong in the lexicon, yet, being constructed out of more than one word, they would also fall within the traditional scope of grammar, even if constituted (cf. *by and large*) from non-standard sequences of syntactic categories. As we shall soon see, the glue that can hold an MWE together often involves grammatical relations between the subparts, so that the structure of linguistic processing tasks such as parsing and machine translation (MT), which depends on a normally clear distinction between word tokens and phrases, has to be re-thought to accommodate MWEs. The issue of MWE handling goes to the heart of natural language processing (NLP) where it raises a number of

1 Theories of grammar such as construction grammar (Fillmore, Kay, and O'Connor 1988) deny a strict distinction between the two and posit a syntax–lexicon continuum.

fundamental problems with a frequency that cannot be ignored. Not surprisingly, it has been and still is a main item on the agenda of several working groups such as the PARSEME network (Savary and Przepiorkowski 2013), of which the authors of this article are members.

The main aim of this survey is to shed light on how MWEs are handled in NLP applications. More particularly, it tries to clarify the nature of *interactions* between MWE processing and downstream applications such as MWE-aware parsing and MT. There is no shortage of proposed approaches for MWE processing and MWE-aware NLP applications. In fact, it is the emergence of approaches in the absence of guiding principles that motivates this article.

There have been other surveys and reviews about MWEs with different scopes. Some concentrate primarily on their linguistic characteristics (Mel'čuk et al. 1999; Calzolari et al. 2002; Sag et al. 2002; Wray 2002). Although this is a valid area of linguistic research, it is not of primary interest to researchers who are addressing the design of computational solutions to the spectrum of problems that MWEs bring into focus. Others are bibliographical reviews/state-of-the-art overviews done in the context of Ph.D. theses (Evert 2005; Pecina 2008) or book chapters (Manning and Schütze 1999; McKeown and Radev 1999; Baldwin and Kim 2010; Seretan 2011; Ramisch 2015), with a narrow scope focusing only on a specific part of MWE processing. In these studies, the subject area is relevant, but the work surveyed tends to be on the micro scale and misses the higher-level insights that one hopes might emerge from the study of such a pervasive phenomenon. Several journal special issues on MWEs (Villavicencio et al. 2005; Rayson et al. 2010; Bond et al. 2013; Ramisch, Villavicencio, and Kordoni 2013) are a showcase for outstanding research in the field, but do not provide a broad overview. In short, a *big picture* is missing from existing reviews and without one it is difficult to compare individual solutions or to reveal that ostensibly different solutions might actually share similar characteristics.

To overcome these shortcomings we felt that it was necessary to create a **conceptual framework** within which both the problems and the different research contributions could be positioned. The goals of this survey are the following:

- Provide a focused overview of how MWEs are handled in NLP applications, thereby focusing on MWE processing rather than MWEs as a linguistic phenomenon,
- Clearly define the task of MWE processing and delineate its two main subtasks, discovery and identification,
- Elaborate on the interaction between MWE processing and two selected NLP use cases, that is, parsing and MT,
- Explain how the key MWE properties, such as discontiguity, variability, and non-compositionality, give rise to challenges and opportunities for MWE processing and MWE-aware applications such as parsing and MT.
- Differentiate previous work according to how MWE processing is timed ("orchestrated") with respect to the underlying application.

1.1 Definitions and Categories

Definitions of MWEs abound (Seretan 2011), driven perhaps by their awkwardness—which causes trouble in many corners of linguistic study—their lack of homogeneity,

and their surprising frequency. The awkwardness arises from the way in which they transcend boundaries imposed by the different subfields of morphology, lexicology, syntax, and semantics. Their lack of homogeneity has led to various categorization schemes that we discuss further subsequently. Definitions observed in the literature are not exactly in disagreement, but tend to stress different aspects according to the identified MWE categories under consideration. Here is an illustrative selection:

- "a multiword unit or a collocation of words that co-occur together statistically more than chance" (Carpuat and Diab 2010)
- "a sequence of words that acts as a single unit at some level of linguistic analysis" (Calzolari et al. 2002)
- "idiosyncratic interpretations that cross word boundaries" (Sag et al. 2002)
- "lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity" (Baldwin and Kim 2010)

The first two focus mainly on the essential *structural* aspects of MWEs evidenced by the unusual co-occurrence of two or more elements within a template of some kind. The complexity of the template can vary widely, from a simple sequence of two fixed words, to longer sequences of less tightly specified elements (e.g., lexemes) constrained by syntactic and/or semantic relationships, with the possibility of intervening gaps.

The third definition emphasizes the essentially idiosyncratic semantic aspect of MWEs, evidenced by degrees of non-compositionality in arriving at the interpretation of the whole from the several parts.

In this article we subscribe to the fourth definition, by Baldwin and Kim (2010), which captures both of these aspects—that is, outstanding co-occurrence (i.e., collocation or statistical idiomaticity) and generalized non-compositionality (i.e., lexical, syntactic, semantic, and pragmatic idiomaticity). This definition also emphasizes that the anomalies of MWEs are manifest over different linguistic levels.

As mentioned earlier, MWEs are not homogeneous and have been categorized using different schemes. The following list defines categories of MWEs commonly seen in the literature. These categories are non-exhaustive and can overlap. They cover the examples mentioned in this article:

- An **idiom** is a group of lexemes whose meaning is established by convention and cannot be deduced from the individual lexemes composing the expression (e.g., *to kick the bucket*).
- A **light-verb construction** is formed by a head verb with light semantics that becomes fully specified when combined with a (directly or indirectly) dependent predicative noun (e.g., *to take a shower*).[2]
- A **verb-particle construction** comprises a verb and a particle, usually a preposition or adverb, which modifies the meaning of the verb and which needs not be immediately adjacent to it (e.g., *to give up*). Verb-particle

2 Light-verb constructions can be called *support-verb constructions* or *complex predicates*, implying slightly distinct notions. We adopt the definition from `http://parsemefr.lif.univ-mrs.fr/guidelines-hypertext/`.

constructions are also referred to as **phrasal verbs** in this article and elsewhere.

- A **compound** is a lexeme formed by the juxtaposition of adjacent lexemes, occasionally with morphological adjustments (e.g., *snowman*).[3] Compounds can be subdivided according to their syntactic function. Thus, **nominal compounds** are headed by a noun (e.g., *dry run*) whereas **noun compounds** and **verb compounds** are concatenations of nouns (e.g., *bank robbery*) or verbs (e.g., *stir fry*). The literature tends to be ambiguous about the necessity of spaces and hyphens for separating the parts of an MWE. Some authors refer to **closed compounds** when they are formed from a single token (e.g., *banknote*), and **open compounds** when they are formed from lexemes separated by spaces or hyphens.
- A **complex function word** is a function word formed by more than one lexeme, encompassing multiword conjunctions (e.g., *as soon as*), prepositions (e.g., *up until*), and adverbials (e.g., *by and large*).
- A **multiword named entity** is a multiword linguistic expression that rigidly designates an entity in the world, typically including persons, organizations, and locations (e.g., *International Business Machines*).
- A **multiword term** is a multiword designation of a general concept in a specific subject field[4] (e.g., *short-term scientific mission*).[5]

For further examples and discussion on the phenomenon itself, the reader is referred to existing surveys (Sag et al. 2002; Baldwin and Kim 2010), workshop proceedings,[6] and Web sites.[7]

1.2 Properties of MWEs

MWEs can be characterized by a number of properties that on the one hand present challenges for MWE processing and the two use cases, namely, parsing and MT, but on the other hand create opportunities for the correct handling of MWEs. We briefly describe the main properties, focusing on those that represent challenges and/or opportunities for NLP applications, in particular, for the NLP applications under consideration in this survey, parsing and MT, that are presented in more detail in the next section (Section 1.3).[8]

Arbitrarily prominent co-occurrence, that is, **collocation**, is one of the outstanding properties of MWEs. For example, although the words *strong*, *powerful*, *intense*, and *vigorous* are (near) synonyms, only *strong* is usually used to magnify the noun *coffee* (Pearce 2001). This property has been heavily used by MWE discovery methods partly because

3 Compounding is a general linguistic phenomenon and not all compounds are MWEs. Compounds can be fully compositional, in which case they are not MWEs (e.g., *paper card*), conventionalized, in which case they are statistically idiomatic MWEs (e.g., *credit card*) or non-compositional MWEs, usually showing some level of semantic idiomaticity (e.g., *green card*).

4 The definition comes from ISO 1087-1:2000.

5 In this article we do not cover single-word named entities and terms, which are by definition not MWEs.

6 The MWE workshop series: `http://multiword.sf.net`.

7 For example: `http://mwe.stanford.edu/`, `http://collocations.de` and `http://parseme.eu`.

8 Other properties often discussed in the literature but not emphasized by this survey include heterogeneity, non-substitutability, lexicalization, and extragrammaticality.

it is easy to capture using statistical **association measures** (Section 2.2.1). Conversely, prominent co-occurrence is problematic for MT, because word-for-word translation might lead to translations of words that are suitable individually, but that yield non-fluent or ambiguous translations of MWEs. For instance, the Italian expression *compilare un modulo* has to be translated into English as *to fill in a form* rather than the word-for-word translation *to compile a module*.

Discontiguity, whereby alien elements can intervene between core MWE components, is a challenge for MWE processing. For instance, the Portuguese expression *levar em conta* (*to take into account*) licenses a direct object that can either appear after the idiom, like in *ele levou em conta minha opinião* (*he took into account my opinion*) or between the verb and the fixed prepositional complement, like in *ele levou minhão opinião em conta* (*he took my opinion into account*). Discriminating the intervening words from the core can be non-trivial but if they form a single syntactic constituent, as in the example, the task can be facilitated by syntactic analysis, thus creating an opportunity for parsing.

Non-compositionality is prototypical in idioms such as the French nominal compound *fleur bleue* (lit. *blue flower*). This expression is used to characterize a sentimental and often naive person, so its meaning is completely opaque to speakers who only know the meanings of the individual words. This property is a challenge for MT because translating non-compositional MWEs through the individual words or structures will very often yield an inappropriate translation. The problem of non-compositionality of MWEs requires a strategy aiming to correctly identify the borders of MWEs and to find the associated sense of the expression. For example, the Romanian idiom *a i-o coace cuiva* (lit. *to bake it for someone*) should be translated in English as *to prepare a trap/prank/ambush*. Several strategies use external resources, often the fruits of MWE discovery, to identify MWEs and their equivalents. Conversely, non-literal translations can serve as a cue for ranking non-compositional MWEs such as idioms during MWE discovery.

Ambiguity is a challenge for many NLP tasks. The type of ambiguity that impacts MWE processing the most is the choice between a compositional and an MWE reading of a sequence of words, as illustrated by the sentence *I am struck by the way the rest of the world is confident of a better future.* In most cases the sequence of words *by the way* is an MWE with the approximate meaning of *incidentally*. However, in the example it is a regular prepositional complement of the verb *struck*. In some cases, syntactic analysis can aid in determining whether the sequence of words should be recognized as an MWE. An analysis that takes *by the way* to be an MWE and thus an adverb in this case, will yield an ungrammatical sentence (which becomes clear when we replace *by the way* with *incidentally*: *I am struck incidentally the rest of the world . . .*). Parsing can help reveal the relevant subcategorization frame that includes the preposition selected by the verb.

Variability, that is, the fact that MWEs allow for varying degrees of flexibility in their formation, poses great challenges for their identification. Searching for fixed forms only will lead to low recall, because the fixed form will fail to match all possible variations. For example, *een graantje meepikken* (lit. *to pick a grain with the others*) is a Dutch MWE meaning to benefit from something as a side effect. Just searching for the fixed string *graantje meepikken* will not identify *Zij pikken er hun graantje van mee* (lit. *they pick their grain of something with the others*), meaning that they are benefiting from something as a side effect. However, syntactic analysis can help us identify the parts of this MWE that allow for variation, here the determiner that can be changed into a possessive pronoun. The problem becomes more serious in morphologically rich languages where words may take hundreds of different surface forms. For example, the Turkish MWE *kafa çekmek* (lit. *to pull head*), referring to consuming alcohol, may appear in

many different surface forms such as *kafaları çekelim* (lit. *let's pull the heads*), *kafayı çektim* (lit. *I pulled the head*), *kafayı çektin* (lit. *You pulled the head*), *kafaları çekecekler* (lit. *They will pull the heads*), among many more under different number and person agreements and tenses.

1.3 Conceptual Framework

The conceptual framework we propose has several features. It offers a clear definition of what we mean by the term MWE processing. It also shows how the inherent properties of MWEs give rise to shared challenges as well as opportunities across MWE processing tasks and use cases, thus revealing the complex pattern of interactions between MWE processing and the two use cases. Finally, it explains how these interactions, and in particular the directions in which they operate, give rise to a number of orchestration scenarios (how MWE processing is scheduled with respect to the use cases).

1.3.1 MWE Processing. MWE processing is composed of two main subtasks that are often confused in the literature: **MWE discovery** and **MWE identification** (as shown in Figure 1). MWE discovery is concerned with finding *new* MWEs (types) in text corpora, and storing them for future use in a repository of some kind such as a lexicon. In contrast, MWE identification is the process of automatically annotating MWEs (tokens) in running text by associating them with known MWEs (types).

The delineation of the two tasks seems fundamental to us because the results of both processes are distinct. The output of discovery is a list of MWE lexical entries, whereas for identification it is a list of annotations. We also distinguish them because they often use different approaches and evaluation strategies. Authors of new discovery methods tend to apply unsupervised techniques that are evaluated in terms of the quality of MWEs discovered. On the other hand, identification approaches are often based on supervised learning models whose results are evaluated by comparing automatically tagged text to reference annotations. An earlier proposal to define MWE processing listed classification and disambiguation as component tasks (Anastasiou et al. 2009). Our framework incorporates both of these concepts, with classification being included in discovery and disambiguation within identification. The two subtasks of MWE processing are further explained in Sections 2 and 3.

We have observed much vagueness and high variability in the definition and nomenclature of these two subtasks in the literature, and hope that this survey will contribute to a more stable terminology in the future, thus easing comparisons between related approaches.

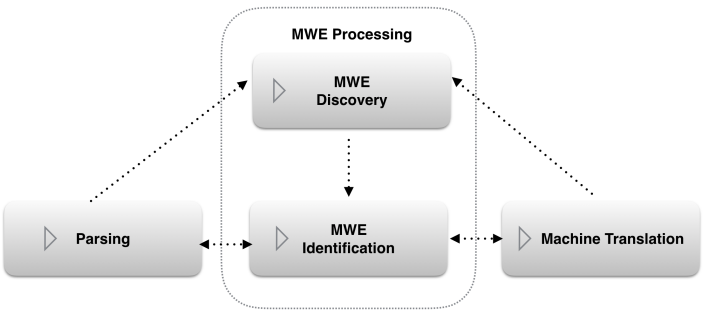


Figure 1
Outline of relations between MWE processing and the two use cases.

1.3.2 Use Cases. The article focuses on two use cases of MWE processing at the heart of language technology for which the correct handling of MWEs is equally crucial: parsing and MT. These uses were chosen because they are representative of past and current efforts to develop **MWE-aware** applications. There have been some attempts to integrate MWE processing into other applications such as information retrieval and sentiment analysis, but we do not cover them in this survey.

Parsing is generally concerned with the definition of algorithms that map strings to grammatical structures. Although MWE-aware parsers represent only a small portion of the total parsing literature, we argue that proper MWE identification can improve parser performance (Cafferkey, Hogan, and van Genabith 2007). In particular, complex function words that have a key role in syntax may be ambiguous (e.g., *by the way*). Failing to identify MWEs will lead to parsing errors. Clearly, a key characteristic of all MWE-aware parsing algorithms is that they *must* in some way have access to pre-existing MWE resources. There are many ways to represent such resources and to incorporate them into the parsing process and this gives rise to the observed variation in the design of such algorithms (Section 4).

MT is more complex than parsing insofar as it involves not only the identification of source MWEs but also their translation into the target language. Although phrase-based approaches aimed at capturing the translation of multiword units and may in principle handle contiguous MWE categories such as compounds, these approaches will certainly not be able to handle discontiguous MWEs, and neither will they cater for variants of MWEs, unseen in the training data. Attempts at MWE-aware MT have shown variable results, according to the category of MWE under consideration and the given language pair, but have proved beneficial in a number of cases (Pal, Naskar, and Bandyopadhyay 2013; Cap et al. 2015). As with parsing, pre-existing resources are necessary and there are several ways to integrate such resources in the translation process (Section 5).

Because the properties of MWEs represent challenges to one process, but opportunities for another (Section 1.2), they induce a complex pattern of bidirectional interactions. Figure 1 gives an overview of the main support relations between the two processes involved in MWE processing and our two selected use cases.

The single arrows in Figure 1 indicate a support relationship. So the arrow from discovery to identification means that discovery supports identification in virtue of the lexical resources that discovery yields. Similarly, the arrows from MT and parsing to discovery indicate that the outputs of both parsing and MT have been shown to support discovery. Syntactic analysis can help deal with discontiguity, as exemplified above, and non-literal translations can serve as a cue for ranking non-compositional MWEs for discovery.[9]

The bidirectional arrows indicate two-way support. Parsing can support identification, for example, when a grammatical relationship must hold between MWE components. Translation can also support identification on the target side given a pair of parallel texts. The converse relations also hold. Identification can support parsing in that the identified MWE legitimates special treatment by the parser. It can also support the correct translation of an MWE identified on the source side.

Note that this picture shows the main support relations found in previous work only. Additional arrows are possible and in Section 6 we argue for a large-scale

9 For MT, the arrow does not mean that the *output* of MT can help discovery, but that *resources* and *tools* used in MT can also be useful for discovery.

evaluation over a systematic set of experiments that cover the less populated areas of the interaction landscape as well.

1.3.3 Orchestration. This complex set of interactions, and in particular the directions in which they operate, give rise to a variety of architectures that differ in how MWE processing is scheduled with respect to the use case. More precisely, they define whether MWE processing is done before (as *preprocessing*), after (as *postprocessing*), or during (*jointly* with) the given use case. Although joint systems perform both tasks simultaneously, preprocessing and postprocessing can be seen as pipelines, in which the output of one process constitutes the input of the next one.

The following sections on MWE discovery, MWE identification, parsing, and MT further explain how the core tasks of MWE processing are incorporated into the use cases and vice versa. In particular, they develop the notion of **orchestration**, the effort of trying to find the best entry-point for one process to help the other—for example, the optimum moment to introduce MWE identification into the parsing pipeline.

The parsing literature reveals that authors have chosen different entry-points for MWE identification in the process. The choice of MWE identification *before* parsing, where methods are used to partly annotate words and sequences in advance, can reduce the search space of the parsing algorithm. Otherwise one can opt to do MWE identification *after* parsing, allowing it to benefit from the available syntactic analysis. MWE identification *during* parsing has the benefit that several alternatives can be maintained and resolved with joint learning models.

We see alternative approaches to orchestration in the literature on MT as well. On the one hand, we find MWE identification *before* translation methods (the so-called static approaches) that concatenate MWEs as a preprocessing step or conversely split compositional closed compounds in Germanic languages to distinguish them from non-compositional compounds. On the other hand, we find MWE identification *during* the translation process itself (so-called dynamic approaches).

1.3.4 Resources. Much of the glue that holds together the network of interactions shown in Figure 1 is composed of resources, which in the case of MWEs fall into three basic categories: lexical resources, (parallel) corpora, and treebanks. **Lexical resources** are essentially databases, and include MWE lexicons and general-purpose dictionaries containing MWE material. Both are useful for handling specific categories of MWE, such as multiword named entities (Steinberger et al. 2013), multiword terms, or idioms. Lexical resources are particularly effective for the identification of highly fixed MWEs. Otherwise, they may be combined with rules describing possible syntactic constraints within MWEs.

Corpora consist of natural text, and may be annotated in different ways. Minimally, tags are simply used to delimit MWEs. Further information, concerning MWE categories, for example, can be added as tag features. Progressively more refined information can approach the level of expressiveness found in treebanks. Examples of annotated corpora with MWE tags include Wiki50 (Vincze, Nagy, and Berend 2011), STREUSLE (Schneider et al. 2014b), and the PARSEME shared task corpora (Savary et al. 2017). Two or more corpora can also be set in correspondence. For example, **parallel corpora** in different languages include sentence-level alignment and are used to detect many-to-many, one-to-many, or many-to-one translations. An example of MWE-annotated parallel corpus is the English–Hungarian SzegedParallelFX corpus (Vincze 2012).

Finally, **treebanks** are special corpora that include *syntactic* relations between nodes over text segments and are arguably the most valuable resources for data-driven parsing

systems and syntax-aware MT systems. In the literature, there exist different opinions on whether syntactically regular but semantically idiomatic MWEs should be identified in syntactic treebanks. Although the Penn Treebank designers prefer not to annotate verbal MWEs (Marcus, Marcinkiewicz, and Santorini 1993), these are annotated in the Prague Treebank (Bejček et al. 2012). For example, whereas the syntactic structure within light-verb constructions such as *to make a decision* is annotated with special MWE relations in the Prague Treebank, they are annotated as regular verb–object pairs in the Penn Treebank. Although, at the time of writing this survey, there is still not a universal standard for MWE annotation, one of the main goals of the PARSEME network is to develop annotation guidelines for MWE representation in both constituency and dependency treebanks. For an up-to-date status of the current annotations for different languages, see Rosén et al. (2015, 2016). Appendix A provides a complementary list of resources and tools for MWE processing.

Identification relies on lexical resources that can be either the fruit of discovery or hand-built. Both parsing and MT rely on lexical resources as well, either through a separate identification step or by using them internally. For example, MWE lexicons are important for MT within preprocessing, postprocessing, and translation phases of different paradigms: They are mainly used to delimit MWEs, replacing them by either a single token, a sense identifier, or by a translation equivalent before alignment takes place. In addition to lexicons, statistical parsing depends on treebanks annotated with MWEs, and MT relies on parallel corpora (or treebanks for syntax-aware MT) to retrieve translation equivalents for the MWEs it has identified.

1.4 Evaluation

The application of MWE identification in parsing and MT opens up possibilities for extrinsic evaluation of the former. By assessing the quality of parsers and MT systems that incorporate various automatic MWE identification methods, and comparing results, the contribution of individual MWE identification methods can be estimated. Similarly, MWE discovery can be evaluated extrinsically by testing the usefulness of (semi-)automatically created MWE lists for identification. The intrinsic evaluation of MWE discovery and MWE identification, on which more details can be found in Sections 2 and 3, is non-trivial, among other things because of the lack of available test data. Evaluating MWE processing extrinsically through the use cases of parsing and MT is an attractive alternative. However, as Sections 4 and 5 will further specify, caution is needed when measuring the impact of MWE identification on, for example, parsing quality. Depending on the type of orchestration at hand, different evaluation strategies should be adopted in order to avoid misleading conclusions.

1.5 Structure of This Article

The survey is organized in six sections as follows. This introduction is followed by two sections that provide a concise overview of approaches to MWE discovery (Section 2) and identification (Section 3), the two subtasks of MWE processing. A transparent definition of MWE processing and a clear delineation of its two subtasks are the result of this. After shedding light on the properties of MWEs that are exploited by MWE approaches and the challenges they face, we are able to understand the interactions between MWE processing and the two use cases, parsing (Section 4) and MT (Section 5). In particular, these sections will show how shared challenges and opportunities give rise

to several possible orchestration scenarios and position previous work with respect to this timing issue. The final section (Section 6) offers conclusions and future perspectives.

2. MWE Discovery

Our survey focuses on interactions of MWE processing with parsing and MT. However, we cannot discuss these interactions without providing an overview of approaches in discovery and identification. Other surveys on these tasks have been previously published (Baldwin and Kim 2010; Seretan 2011; Ramisch 2015). Our main contributions are to cover the latest advances and group references across languages and MWE categories according to each method's characteristics. Hence, the goal of this section is to define MWE discovery and provide a concise overview of the current state of affairs.

This section describes existing approaches for MWE discovery. As defined in Section 1.3, discovery is a process that takes as input a text and generates a list of MWE candidates, which can be further filtered by human experts before their integration into lexical resources. This process is depicted in Figure 2.

Automatic MWE discovery (hereafter simply referred to as *discovery*) has been an active research topic since the end of the 1980s when a number of seminal papers were published (Choueka 1988; Church and Hanks 1990). The famous "pain-in-the-neck" paper (Sag et al. 2002) and the related MWE workshops (Bond et al. 2003) have put discovery in focus as one of the main bottlenecks of NLP technology. Since then, considerable progress has been made, notably in the context of national and international research projects like PARSEME (Savary et al. 2015).

Whereas discovery methods generate lists of MWE types out of context, MWE identification marks MWE tokens in running text. However, several terms have been used to designate what we have defined as *discovery* in our conceptual framework (Section 1.3), such as *identification*, *extraction*, *acquisition*, *dictionary induction*, and *learning*. Because one of the aims of this article is to clearly delineate the tasks of, on the one hand, discovering MWE types, and on the other, identifying MWE tokens in running text (Section 3), discovery seemed the most suitable term at the right level of specificity. Our survey focuses on *empirical* strategies for MWE discovery as opposed to *expert lexicon construction* by human language experts. *Empirical* methods try to automatically learn lexical information from textual data. In practice, empirical and expert methods

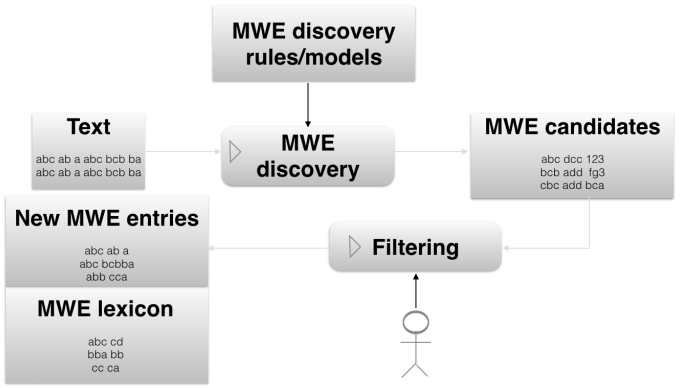


Figure 2
MWE discovery: extract MWE information from corpora to create or enrich lexicons.

are complementary and both can be combined when one builds a lexical resource, as indicated by the human intervention in Figure 2.

Most discovery methods leverage characteristics of MWEs (Section 2.1) to design measures, rules, and heuristics for discovering new MWEs in data. For example, collocation strength is exploited by methods measuring the association between tokens to distinguish MWEs from regular phrases. In addition, supervised machine learning relies on annotated data, using linguistic knowledge for feature engineering.

2.1 Motivations and Challenges for MWE Discovery

Because of the prevalence and productivity of MWEs, discovery methods are indispensable. They are used to speed up expert lexicon creation or to update existing lexicons with novel entries that are created constantly. Robust NLP systems rely on unsupervised or semi-supervised methods to discover new expressions as they are coined or become popular. Discovery enables identification, as shown by the arrow going from discovery to identification in Figure 1. In other words, discovery helps extending and creating lexicons that are used in identification. However, there are also some challenges in MWE discovery, the most relevant of which are *discontiguity* and *variability*.

Discontiguity. Discovering flexible constructions, like verbal expressions, is a challenge because of their discontinuous nature. A discovery method must be able to locate all the elements of the MWE even if they are separated by arbitrary intervening material. Discontiguity is generally dealt with when parsing the input corpus prior to MWE discovery (Section 4.1). However, syntactic analysis may introduce ambiguities because some MWEs do not allow arbitrary modification and will be wrongly merged with literal uses or accidental cooccurrence (*to take turns* does not mean that one deviates from a straight trajectory several times. It means that two or more people do something alternately).[10]

Variability. Especially in morphologically rich languages, variability poses problems at various levels. Returning several variants of the same MWE slows down manual filtering as the same MWE is presented several times to lexicographers. For empirical methods, variability increases data sparsity. For example, conflating all variants into a single MWE candidate results in more reliable association estimators. Variability can also be partly addressed by linguistic analysis. For example, variant conflation often requires lemmatization in order to merge inflections. Nonetheless, some problems remain because automatic POS taggers, lemmatizers, and parsers are not perfect. As a consequence, current discovery methods generally yield rough material for lexicon construction that generally requires a good deal of polishing before becoming useful.

2.2 Main MWE Discovery Methods

Some methods are designed to deal with specific categories of MWEs, for instance, focusing on noun compounds (Girju et al. 2005; Salehi, Cook, and Baldwin 2015), light-verb constructions (Stevenson, Fazly, and North 2004), or verb-particle constructions (McCarthy, Keller, and Carroll 2003; Ramisch et al. 2008b). Others are generic and deal uniformly with many MWE categories (da Silva et al. 1999; Seretan 2011). In any case,

10 Whereas *to take turns* means *to do something alternately*, *to take a turn* means simply *to turn, to deviate from a straight trajectory*.

discovery methods can be differentiated according to the linguistic properties of MWEs that they leverage, some of which have already been discussed (Section 1.2).

- **Collocation**. Words that are part of an MWE tend to co-occur more often than expected by chance. This property is generally modeled by methods based on association measures such as prominent co-occurrence (Section 2.2.1).
- **Non-substitutability**. It is not possible to replace part of an expression by a synonym or similar word. This property is generally modeled by variability or fixedness measures (Section 2.2.2).
- **Non-compositionality**. The meaning of the whole expression cannot be inferred from the meanings of its parts. This property is generally leveraged by models based on vector-space semantic similarity (Section 2.2.3).
- **Non-literal translatability**. Word-for-word translation tends to generate unnatural, ungrammatical and sometimes nonsensical results. Monolingual and multilingual discovery methods based on translation asymmetries use techniques inspired from MT, and are thus presented later (Section 5.2.1).

These properties are not orthogonal. For example, non-literal translation and non-substitutability are side-effects of non-compositionality, and because non-compositionality is hard to model, such derived properties are additionally used in discovery.

The use of **morphosyntactic and syntactic patterns** is quite common to generate a first list of **MWE candidates**[11] of a specific category. For instance, a list of candidate nominal compounds in English can be obtained by looking for nouns preceded by other nouns or adjectives. Justeson and Katz (1995) suggest a limited set of seven bigram and trigram POS patterns, combining nouns and adjectives, in order to discover nominal compound candidates in English. Baldwin (2005) investigates the use of increasingly sophisticated morphosyntactic and syntactic patterns to discover new verb-particle constructions in English.

Because such patterns are used as a preprocessing technique by many discovery strategies presented in the remainder of this section, we do not discuss them in detail. The benefits of using POS-taggers and parsers to help discovery is represented by the unidirectional arrow from parsing to discovery in Figure 1 and will be discussed further in Section 4.2.1.

2.2.1 Association Measures. Prominent co-occurrence, that is, collocation, is the simplest MWE property to model in computational systems. It can be accurately captured by statistical **association measures**, which estimate the association strength between words in a corpus based on their co-occurrence count and on their individual word counts. Most measures take into account the observed co-occurrence count of a group of n words w_1, w_2 $\ldots w_n$ compared with its expected count. The expected co-occurrence

11 An MWE candidate is a sequence or group of words in the corpus that match a given pattern.

count is based on the assumption that words are independent, that is, it equals the product of their individual word probabilities.[12]

A popular association measure in MWE discovery is **pointwise mutual information**. It was first proposed in terminology discovery by Church and Hanks (1990), and can be expressed as the log-ratio between observed and expected counts. Values close to zero indicate independence and the candidate words are discarded, whereas large values indicate probable MWEs. Other measures are based on hypothesis testing. If we assume as null hypothesis that words are independent, their observed and expected counts should be identical. Using a test statistic like Student's t, large values are strong evidence to reject the independence null hypothesis, that is, the candidate words are not independent and probably form an MWE.

More sophisticated test statistics for two-word MWE candidates take into account their *contingency table*. Examples of such measures are χ^2 and the more robust likelihood ratio (Dunning 1993). Pedersen (1996) suggests using Fisher's exact test in automatic MWE discovery, and this measure is implemented among others in the Text:NSP package.[13] Another measure for MWE discovery is the average and standard deviation of the distance between words, implemented in Xtract (Smadja 1993). Because these measures are based on frequency counts, there have been some studies to use Web hits as an alternative to corpus counts, in order to avoid low-frequency estimates (Keller and Lapata 2003; Ramisch et al. 2008a).

Although association measures work quite well for two-word expressions, they are hard to generalize to arbitrary *n*-word MWE candidates. One simple approach is to merge two-word MWEs as single tokens and then apply the measure recursively. For instance, in French, the MWE *faire un faux pas* (lit. *to make a false step*, 'to make a blunder') can be modeled as the verb *faire* (*to make*) combined with the compound *faux_pas* (*blunder*), which had been merged due to high association in a previous pass (Seretan 2011). The LocalMaxs algorithm finds optimal MWE boundaries by recursively including left and right context words, stopping when the association decreases (da Silva et al. 1999).[14] A similar approach, using a lexical tightness measure, was proposed to segment Chinese MWEs (Ren et al. 2009).

Association measures can be adapted according to the morphosyntactic nature of lexical elements. Hoang, Kim, and Kan (2009) propose new measures where very frequent words such as prepositions are weighted differently from regular tokens. Comparisons between different association measures have been published, but to date no single best measure has been identified (Pearce 2002; Evert 2005; Pecina 2008; Ramisch, De Araujo, and Villavicencio 2012).

2.2.2 Substitution and Insertion. A *French kiss* cannot be referred to as a *kiss that is French*, a *kiss from France*, or a *French smack*, unlike non-MWE combinations like *French painter* and *passionate kiss*. Because of their non-compositionality, MWEs exhibit **non-substitutability**, that is, limited morphosyntactic and semantic variability. Thus, the replacement or modification of individual words of an MWE often results in unpredictable meaning shifts or invalid combinations. This property is the basis of discovery methods based on substitution and insertion (including permutation, syntactic alternations, etc.).

12 For more details on association measures, see `http://www.collocations.de`, Evert (2005), and Pecina (2008).

13 `http://search.cpan.org/dist/Text-NSP/`.

14 `http://research.variancia.com/multiwords2/`.

Pearce's (2001) early synonym substitution method replaces parts of the MWE by synonyms obtained from WordNet, and then obtains frequencies for the artificially generated MWE variants from external sources. Instead of using variant frequencies directly, it is possible to estimate an MWE candidate's frequency using a weighted sum of variant corpus frequencies (Lapata and Lascarides 2003) or Web-based frequencies (Keller and Lapata 2003). A similar approach is used by Villavicencio et al. (2007) and Ramisch et al. (2008a), but instead of synonym variations, the authors generate syntactic permutations by reordering words inside the MWE, combining frequencies using an entropy measure. Artificially generated variants can be transformed into features for supervised discovery methods, as we will see in Section 2.2.4 (Lapata and Lascarides 2003; Ramisch et al. 2008a).

Methods based on variant generation and/or lookup were used to discover several MWE categories, such as English verb-particle constructions (McCarthy, Keller, and Carroll 2003; Ramisch et al. 2008b), English verb-noun idioms (Fazly and Stevenson 2006; Cook, Fazly, and Stevenson 2007), English noun compounds (Farahmand and Henderson 2016), and German noun-verb and noun-PP idioms (Weller and Heid 2010).

Such methods often require external lexicons or grammars describing possible variants, like synonym lists or local reorderings (e.g., $Noun_1\ Noun_2 \rightarrow Noun_2\ of\ Noun_1$). Synonyms or related words in substitution methods can come from thesauri like a WordNet and VerbNet (Pearce 2001; Ramisch et al. 2008b). Related words can be found in automatically compiled thesauri built using distributional vectors (Riedl and Biemann 2015; Farahmand and Henderson 2016). When compared with association measures, most of these methods are hard to generalize, as they model specific limitations that depend on the language and MWE category.

2.2.3 Semantic Similarity. Models based on *semantics* account for the fact that many MWE categories are partly or fully **non-compositional**. Because the meaning of the parts does not add up to the meaning of the whole, there should be little similarity between the computational-semantic representation of MWEs and of words that constitute them. For instance, let us consider the items *cat*, *dog*, *hot dog*, and *sandwich*. We would expect that *dog* is similar to *cat*, *dog* is not similar to *hot dog*, and *hot dog* is similar to *sandwich*.

Semantic similarity methods differ mainly in how they represent word and MWE senses, how they combine senses, and how they measure similarity. Word and MWE senses can be modeled using entries of semantic lexicons like WordNet synsets (McCarthy, Venkatapathy, and Joshi 2007). However, most discovery methods use distributional models (or word embeddings) instead, where senses are represented as vectors of co-occurring context words (Baldwin et al. 2003; Korkontzelos 2011). The creation of such vectors in distributional models has several parameters that affect the performance of MWE discovery, such as the number of vector dimensions and the type of context window (Cordeiro et al. 2016). The evaluation of discovery methods based on distributional similarity can use dedicated test sets (Reddy, McCarthy, and Manandhar 2011; Farahmand and Henderson 2016) or use handbuilt resources such as WordNet (Baldwin et al. 2003).

Because methods based on distributional semantics use contextual information to represent meaning, they are closely related to substitution methods described in Section 2.2.2. For instance, Riedl and Biemann (2015) design a measure that takes into account the similarity of an MWE with single words that appear in similar contexts. They assume that MWEs tend to represent more succinct concepts, thus their closest

distributional neighbors tend to be single words. Therefore, their method can be classified both as a substitution-based method (replace an MWE by a single word in context) and as a semantic similarity method.

There are several ways to combine and compare distributional vectors for MWE discovery. McCarthy, Keller, and Carroll (2003) consider the overlap between the set of distributional neighbors of a verb-particle construction and its single-verb counterpart. For instance, if *to break up* and *to break* share many neighbors, then *to break up* must be more compositional than *to give up*, which has no shared neighbor with *to give* (Baldwin et al. 2003). A popular measure to discover idiomatic MWEs is the cosine similarity between the MWE vector and the member word vectors (Baldwin et al. 2003; Reddy, McCarthy, and Manandhar 2011). Salehi, Cook, and Baldwin (2015) compare two similarity measures: (a) the weighted average similarity of the MWE vector with its member word vectors, and (b) the similarity between the MWE vector and the average vector of the component words.

Semantic similarity methods have been successfully evaluated on small samples of verb-particle constructions (Baldwin et al. 2003; Bannard 2005), verb-noun idioms (McCarthy, Venkatapathy, and Joshi 2007), and noun compounds (Reddy, McCarthy, and Manandhar 2011; Yazdani, Farahmand, and Henderson 2015; Cordeiro et al. 2016). Their adaptation to large-scale discovery remains to be demonstrated.

2.2.4 Supervised Learning. Supervised learning approaches for discovery use annotated data sets as training material to learn how to distinguish regular word combinations from MWEs. Often, the features used in supervised methods include scores derived from unsupervised methods discussed above, such as association measures and semantic similarity.

The use of these as features has proven to be an effective way to combine scores, giving more weight to more discriminating features and reducing the weight of redundant ones (Ramisch et al. 2008a). It also provides a workaround for the problem of choosing a scoring method for a given data set among dozens of methods proposed in the literature. Furthermore, the learned models can provide insight into features' informativeness (Ramisch et al. 2008b).

One of the first experiments using a supervised approach was proposed by Lapata and Lascarides (2003). The authors use a C4.5 decision tree to classify noun-noun compounds into true MWEs and random co-occurrence. Logistic regression, linear discriminant analysis, support vector machines, and neural networks have been used as classifiers for collocation discovery in Czech and German (Pecina 2008). Rondon, Caseli, and Ramisch (2015) propose an iterative method for the perpetual discovery of novel MWEs. The system requires some initial supervision to build a seed MWE lexicon and classifier, and incrementally enriches it by mining texts in the web and bootstrapping from its results.

Yazdani, Farahmand, and Henderson (2015) use light supervision in the form of a list of noun compounds automatically extracted from Wikipedia. They are used as training material to tune their composition function parameters. A similar approach was also used by Farahmand and Henderson (2016) to model MWE substitutability. Supervised methods are generally very precise but cannot be systematically preferred, as they require annotated data sets. Unfortunately, such data sets are usually (1) not readily available, (2) quite small and specific, and (3) not applicable when the target MWEs are highly ambiguous.

2.3 Evaluation of MWE Discovery

Discovery is a process whose goal is to find *new* lexical entries. Its evaluation is tricky because the utility and relevance of these entries is hard to assess. Most discovery methods output ranked MWE lists, which can be evaluated as follows:

- **Post hoc human judgments**: Given the top n MWEs retrieved by a method, experts select the relevant ones. The proportion of positive entries found is used to evaluate the method (da Silva et al. 1999; Seretan 2011).
- **Dictionaries**: The returned list can be automatically evaluated by checking the entries already present in a gold standard dictionary. This assumes that entries absent from the dictionary are wrong (Ramisch, De Araujo, and Villavicencio 2012; Riedl and Biemann 2015).
- **Dedicated data sets**: Given a list of known positive and negative MWE examples, true MWEs should be ranked first. This can be seen as an information retrieval problem, and measures like average precision, precision at k, recall, and F1 can be used to evaluate the discovery method (Evert 2009; Pecina 2008; Yazdani, Farahmand, and Henderson 2015).
- **Extrinsic evaluation**: Because evaluating discovered lexical entries is tricky, we can evaluate the performance of downstream tasks that use them, such as identification (Riedl and Biemann 2016), parsing (Villavicencio et al. 2007), and MT (Costa-Jussà, Daudaravicius, and Banchs 2010). In this case, not only discovery but also integration into the target application is evaluated. Noise in discovery's output often prevents integrated systems from showing significant performance gains. Extrinsic evaluation is further discussed in the following sections.

2.4 Open Issues in MWE Discovery

Association measures work well with two-word MWEs, but not straightforwardly for longer chunks, especially for measures relying on contingency tables (Banerjee and Pedersen 2003). Nesting and variability also pose problems for their use. Moreover, results are not systematically corroborated and the success of an association measure seems to be heavily dependent on corpus size, nature, and MWE category (Dunning 1993; Pearce 2002; Evert 2005; Pecina 2010).

Substitution and semantic similarity methods require large corpora with many occurrences of the target MWEs. In this kind of approach, discovery is often modeled as a task that consists of finding non-compositional combinations among a list of candidate MWEs. Therefore, their evaluation often requires data sets annotated with semantic compositionality, which are not easy to build. Even though such resources do exist, they are rare and often quite small, mainly available for English, with a few exceptions including nominal compounds in German (Schulte im Walde, Müller, and Roller 2013) and French (Cordeiro et al. 2016).

The main open issue concerns the trade-off between evaluation and usefulness. On the one hand, evaluation sets are not always available and, when they are, results may be hard to generalize because of their small size. On the other hand, large-scale discovery, although potentially useful, is hard to evaluate. As a result, methods tend to be specialized and will not work so well when ported to other MWE categories and

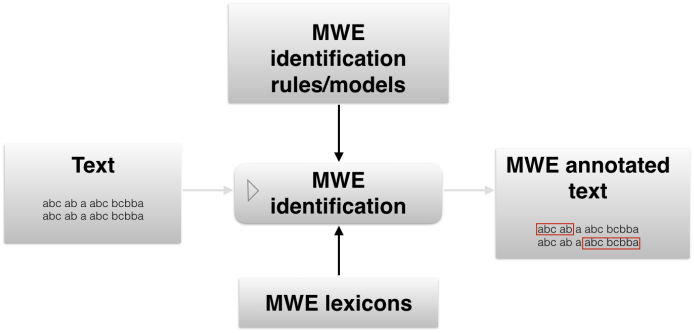


Figure 3
MWE identification: find occurrences of MWEs in running text.

languages. Finally, there has been little work on placing discovery methods back into the bigger picture—in other words, comparing discovery with manual lexicon construction. Therefore, it remains unclear whether discovery is required or even useful to support and/or replace lexicographic work in production scenarios.

One promising alternative is the extrinsic evaluation of discovery to help downstream tasks. Although some authors show that discovery can help MT quality (Costa-Jussà, Daudaravicius, and Banchs 2010), research on identification and MWE-aware parsing strongly relies on handcrafted lexicons (Schneider et al. 2014a; Constant and Nivre 2016) and rarely uses discovery results (Riedl and Biemann 2016). This is often due to a mismatch between discovered MWEs and test sets, and to the difficulty in integrating such noisy lexicons into applications. As a consequence, it is rare to observe convincing performance gains in downstream tasks. We believe that future research should focus on developing extrinsic evaluation measures, test sets, and guidelines for MWE discovery.

3. MWE Identification

The last section reviewed previous work in discovery, the first subtask of MWE processing as defined in the framework presented in Section 1.3. We now turn to MWE identification, the second subtask that, together with discovery, constitutes MWE processing. It has often been considered as a preprocessing step for parsing and MT systems. Many NLP systems perform some sort of MWE identification using a lexicon and direct string matching. This, is often not enough however (Section 3.1). MWE identification takes a corpus as input and adds a layer of annotation indicating where MWE instances occur. The identification process sometimes requires additional input, in the form of MWE lexicons and rules (or models) for detecting MWE instances, as shown in Figure 3. These lexicons may result from automatic MWE discovery (see Figure 1). Identification is in some ways similar to named entity recognition, but MWEs must have at least two words,[15] and some MWEs can be discontiguous.

A system that performs automatic MWE identification is often called an **MWE tagger**. Suppose the input of an MWE tagger is the first row of Figure 4 (which for

15 Note that some MWEs are formed by two or more words but form only one token, such as closed compounds.

Now	**that**	I	**looked**	the	**dirty**	**word**	**up**	,	I	understand	.
1	1		2		3	3	2				
B	I	O	B	o	b	i	I	O	O	O	O

Figure 4
Example of sentence in which MWEs are identified (in **bold**) - and corresponding IOB encoding after Schneider et al. (2014a). The first token of an MWE is tagged with uppercase `B`, the following ones with uppercase `I`, uppercase `O` marks tokens not belonging to any MWE, lowercase `o` indicates non-MWE tokens inside MWE gaps, and lowercase `b` and `i` are used for embedded MWEs. The second row shows token-based MWE identifiers, after Savary et al. (2017).

clarity omits output from initial annotation layers like automatic POS tags or syntactic trees). The expected output is an annotation layer indicating which tokens are part of MWEs (words in bold). The second row provides identifiers that distinguish different MWEs occurring in the same sentence. The MWE tagger might also provide some sort of classification, for example, indicating that the first MWE is a *complex conjunction*, the second one is a *verb-particle construction*, and so on. However, to obtain such output, certain challenges must be addressed.

3.1 Motivations and Challenges for MWE Identification

Identification is important for many NLP applications including both parsing and MT. A parser will have to deal with less ambiguity if some MWEs are identified. For instance, the word *green* has two possible attachments in *green card office*, but if we identify *green card* as a single unit, the ambiguity is eliminated. An MT system can also benefit from MWE identification since many MWEs have no word-for-word translation. For instance, translating *green card* as a phrase prevents invalid reordering like *green office for cards*.

More generally, identification benefits other tasks involving semantic processing. Semantic parsing and role-labeling systems can consider light-verb constructions and idioms to build predicate-argument structures (Bonial et al. 2014; Jagfeld and van der Plas 2015). For example, in the light-verb construction *to take care*, the noun *care* expresses the actual semantic predicate, whereas the verb only links the subject with the nominal predicate (it is semantically "light"). Whereas a syntactic parser should consider that *care* is a syntactic argument of *to take*, a semantic parser predicting argument structure would preferably identify the noun *care* as the predicate rather than considering the verb *to take* as a predicate, with *care* as one of its semantic arguments. Information retrieval systems can use MWEs as indexing keyphrases. Word sense disambiguation systems can avoid assigning spurious labels to individual words of MWEs (e.g., *dry* in *dry run*). The issue of what stage to introduce identification into the processing pipeline of these applications leads to several orchestration scenarios that are discussed in detail in Sections 4.2 and 5.2.

Discontiguity. One challenge in identification is posed by **discontiguous** occurrences such as *look up* in Figure 4. This is particularly important for flexible expressions: Verbal MWEs whose fixed arguments allow reordering or non-verbal MWEs with open slots allowing the insertion of variable components. The overall impact of discontiguity is language-dependent. For example, separable verb-particle constructions, frequent in Germanic languages, are almost non-existent in Romance languages.

Overlaps. A discontiguous MWE can have other nested MWEs in between its components, like *dirty word* as the direct object of *look up*. This is especially problematic for systems that use **IOB encoding**, as exemplified in row 3, which addresses the segmentation problem with tags `B`, for begin, `I` for inside, `O`, for outside (Ramshaw and Marcus 1995). Often, nesting demands multi-level tags, otherwise different segments could be mixed up. For instance, *word* and *up* would have subsequent `I` tags, although they are not part of the same MWE. In the example, outer MWEs use capital `IOB` tags and inner MWEs use lowercase `iob` tags, following the tagging scheme proposed by Schneider et al. (2014a).

Nesting is a particular case of overlap, whereby MWEs can also share tokens in a sentence. For instance, the verb-particle construction *to let out* can be contained in the idiom *to let the cat out of the bag*. If the MWE tagger cannot output more than one MWE identifier per token, it cannot model overlap. One possible workaround would be always choosing the longest sequence. However, this will not work if MWEs share several tokens, but one MWE is not a factor of the other one despite sharing some elements, as in coordinated structures such as *He took a walk and a shower*.

Ambiguity. Even if an MWE tagger has access to a lexicon containing the MWEs *now that*, *look up*, and *dirty word*, it may be totally plausible to use these expressions with their literal meaning: for example, *I realize now that he never looks up at me when I speak*. The use of parsers often helps solve these ambiguities (Section 4.1).

Variability. Some expressions have variable parts or impose non-lexicalized constraints on other elements of the sentence. It might be hard to distinguish elements that are part of an MWE from those that are not. For example, in *have one's say* and *give oneself up*, it is unclear whether the possessive and reflexive clitics are part of the MWE. Morphological analysis and parsing may help identify canonical forms and normalize word order (Section 4.1).

3.2 Main MWE Identification Methods

We distinguish four techniques used in the literature for MWE identification. **Rule-based methods** apply rules of various levels of sophistication to project MWE lexicons onto corpora (Section 3.2.1). **Classifiers** typically used for word sense disambiguation can be adapted to token-based MWE classification using contextual features (Section 3.2.2). **Sequence tagging models**, inspired by POS-tagging, chunking, and named entity recognition, can be learned from manually annotated corpora using supervised techniques (Section 3.2.3). Identification can also be performed as a by-product of *parsing*, as discussed further in Section 4.

3.2.1 Rule-Based Methods. Rules can be as simple as direct matching, but can also use more sophisticated, context-sensitive constraints encoded as finite-state transducers for instance. Historically, rules based on finite-state transducers offered a simple generic framework to deal with variability, discontiguity, and ambiguity (Gross 1989; Breidt, Segond, and Valetto 1996). Identification methods for contiguous MWEs such as open compounds are generally based on dictionaries compiled into finite-state transducers. Standard pattern matching algorithms for finite-state transducers are then applied to the text to identify MWEs. These approaches have two advantages: Dictionaries are compressed using minimization algorithms and matching is extremely efficient in terms of time complexity.

Nonetheless, all inflected forms of MWEs are listed and usually dictionaries are automatically generated from lists of MWE canonical forms together with inflection

rules. An example of this approach is integrated in the MT system Apertium (Forcada et al. 2011). The variable part of MWE lexical entries, as the verb *echar* in the Spanish expression *echar de menos* (lit. *throw from less*, 'to miss'), is inflected according to a regular verbal inflection paradigm. The inflection process may be based on finite-state transducers as in Silberztein (1997), possibly augmented with a unification mechanism for handling agreement between the MWE components (Savary 2009). These approaches are extremely precise, but costly. The manual assignment of inflection rules may be eased by tools like Leximir for predicting inflection classes (Krstev et al. 2013).

Another approach comprises two processing stages: morphological analysis of simple words followed by a composition of regular rules to identify MWEs, as in Oflazer, Çetinoğlu, and Say (2004) for Turkish. Breidt, Segond, and Valetto (1996) design regular rules that handle morphological variations and restrictions like the French idiom *perdre ADV* :la :tête* (lit. *lose ADV* :the :head*, 'to lose one's mind'),[16] lexical and structural variations (*birth date* = *date of birth*). Copestake et al. (2002) design an MWE lexicon for English based on typed feature structures that may rely on analysis of internal words of MWE. Silberztein (1997) also proposes the use of local grammars in the form of equivalence graphs. These approaches are very efficient in dealing with variability and short-distance discontiguity.

Constraints encoded in the lexicon, such as obligatory or forbidden transformations, can be projected on text to disambiguate idiomatic constructions. Hashimoto, Sato, and Utsuro (2006) encode in a lexicon detailed properties of 100 Japanese verb-noun ambiguous idioms such as voice, adnominal modifications, modality, and selectional restrictions. Then, they only classify as idioms those occurrences that match the constraints in a dependency-parsed test set.

More recent approaches to rule-based identification use dictionaries containing canonical MWE forms with no additional constraints. They consist of two stages: (1) POS tagging and lemmatizing the text and (2) performing dictionary lookup (Carpuat and Diab 2010; Ghoneim and Diab 2013). The lookup relies on a maximum forward matching algorithm that locates the longest matching MWE. This simple method handles morphological variants of the MWEs, but tends to overgenerate them. This overgeneration is due to strong morphological constraints on some elements or agreement. For instance, the French idiom *prendre la porte* (lit. take the door) meaning *get sacked* has a strong morphological constraint: the noun *porte* (door) must be in the singular; if it is in the plural, the sequence has its literal meaning. Therefore, using lemmas to identify variants is a potential source of mistakes.

To handle discontiguity, it is possible to apply patterns on such preprocessed texts, including wildcards. For instance, Ramisch, Besacier, and Kobzar (2013) identify discontiguous verb-particle constructions in English made of a verb + at most five words + a particle, adapting the discovery method proposed by Baldwin (2005). General tools for deterministic MWE annotation like the mwetoolkit (Ramisch 2015)[17] allow fine tuning of matching heuristics, minimum and maximum gap size, surrounding POS patterns, and local variants. For example, it is possible to identify verb-particle constructions in which the particles *up* or *down* appear not further than five words after a content verb, constraining intervening words not to be verbs and the next word not to be the word *there* (to avoid regular verbs followed by *up/down there*). The corresponding

16 Colon prefix stands for invariable word; ADV stands for adverbial, asterisk stands for repetition.
17 `http://mwetoolkit.sf.net`.

multi-level regular expression in the mwetoolkit syntax would be [pos∼/VV.*/] [pos!∼/V.*/]{repeat={0,5}} [lemma∼/(up|down)/] [lemma!=there].[18]

Some rule-based identification approaches output ambiguous MWE segmentation, postponing disambiguation until more linguistic context is available (Chanod and Tapanainen 1996). The MWE identification process often generates acyclic finite-state automata representing all possible segmentations for a given input sentence. Some finite state preprocessing tools allow ambiguous lexical analyses, like Intex, Macaon, Nooj, SxPipe, and Unitex.[19] This approach can be used as a preprocessing stage of MWE-aware parsing (Section 4) and as a source of features for sequence-tagging identification (Section 3.2.3).

3.2.2 Sense Disambiguation Methods. Methods inspired by word sense disambiguation treat MWE identification as a specialized in-context classification task. Given a candidate combination in context, the classifier must decide whether it is a true MWE or just a regular co-occurrence. Common features for this task include surrounding words, their POS, lemmas, syntactic characteristics, and distributional information. Such methods often do not cover the identification of candidates. They assume that another process pre-identifies potentially idiosyncratic combinations, and instead focus on detecting which of these are true MWEs. We discuss both supervised and unsupervised classifiers.

Uchiyama, Baldwin, and Ishizaki (2005) tackle the problem of identifying Japanese verb compounds. Sense labels correspond to the meaning added by the second verb (aspectual, spatial, adverbial) with respect to the first verb. Their support vector machine guesses the possible semantic classes of a given verb combination, using the semantic classes of other co-occurring verbs as features. Then, in a second step, identification proper is done simply by taking the most frequent sense.

Hashimoto and Kawahara (2008) propose a supervised disambiguation system able to distinguish literal from idiomatic uses of Japanese idioms. Fothergill and Baldwin (2012) perform an extended evaluation using the same data set and methodology, including new features, a feature ablation study, and cross-idiom tests. Similar approaches based on support vector machines and surface-level features have also been proposed for English light-verb constructions and verb-particle constructions (Tu 2012).

Birke and Sarkar (2006) present a nearly unsupervised system capable of distinguishing literal from non-literal verb uses. It uses a clustering strategy that tries to maximize transitive similarity with the seed set of literal or non-literal sentences using standard features. Sporleder and Li (2009) propose a completely unsupervised method based on lexical chains and text cohesion graphs. Their classifier considers an expression as literal if its presence in the sentence does not have a negative impact on cohesion, defined as the similarity between co-occurring words. For instance, *play with fire* reinforces cohesion in a sentence containing *grilling*, *coals*, and *cooking* but it would reduce lexical cohesion in a chain containing *diplomacy*, *minister*, and *accuse*.

18 This pattern describes a sequence of lexical items. Every lexical item is defined between square brackets by using different linguistic features (e.g., *lemma*, *pos*). The feature value can be set using exact string matching (operator =) or regular expression matching (operator ∼), and both can be negated (operator !). For instance, [pos!∼/V.*/] indicates any item whose part of speech is not prefixed by *V*; [lemma!=there] indicates any item whose lemma is not the string *there*. Every lexical item occurs once by default. In case of potential multiple occurrences, it is followed by a *repeat* feature (between curly brackets) indicating the number of times it can occur as a range (here, [pos!∼/V.*/] can occur between 0 and 5 times).

19 `http://intex.univ-fcomte.fr`, `http://macaon.lif.univ-mrs.fr`, `http://www.nooj4nlp.net`, `http://alpage.inria.fr/~sagot/sxpipe.html`, `http://www-igm.univ-mlv.fr/~unitex/`.

Katz and Giesbrecht (2006), detecting idiomatic verb-noun expressions in German, assume that the context of an idiomatic MWE differs from the contexts of its literal uses. Given two distributional vectors representing literal and idiomatic instances, a test instance is classified according to its similarity to the respective vectors. Cook, Fazly, and Stevenson (2007) propose a similar method based on canonical forms learned automatically from large corpora. Once a canonical form is recognized, distributional vectors for canonical and non-canonical forms are learned and then an instance is classified as idiomatic if it is closer to the canonical form vectors.

Boukobza and Rappoport (2009) mix the supervised and unsupervised approach of Fazly, Cook, and Stevenson (2009) into a single supervised method to identify English verbal MWEs. In addition to literal and idiomatic uses, they also discuss and model accidental co-occurrence in which the member words of an MWE appear together only by chance, proposing the use of specialized multi-way support vector machines for each target candidate. Besides surface contextual features, they exploit the use of automatically obtained syntactic dependencies as features, which sometimes improves precision.

3.2.3 Sequence Tagging. It is possible to build MWE taggers from annotated corpora using supervised structured learning. MWE taggers that model identification as a tagging problem use stochastic models such as conditional random fields, structured perceptron, or structured support vector machines combined with an `IOB` labeling scheme (see Figure 4) to predict MWE labels, given the local context, token-level features, and sometimes external resources.

Blunsom and Baldwin (2006), whose work's main purpose is to acquire new tagged lexical items for two **head-driven phrase structure** grammars (ERG for English and JACY for Japanese), propose a supertagging approach based on conditional random fields to assign lexical types to the tokens of an input sequence using a pseudo-likelihood method that accommodates large tag sets. The proposed approach only enables the identification of contiguous MWEs.

This work is very close to methods for joint (contiguous) MWE identification and POS tagging based on linear conditional random fields (Constant and Sigogne 2011; Shigeto et al. 2013). Their tagging scheme concatenates lexical segmentation information (B and I tags for the IOB tag set) with the POS tag of the lexical unit to which the current token belongs. Constant and Sigogne (2011) trained and evaluated their models on the French Treebank, and Shigeto et al. (2013) worked on a modified version of the Penn Treebank onto which complex function words from Wiktionary were projected.

Some MWE taggers concentrate on the identification task only, using an IOB-like annotation scheme (Vincze, Nagy, and Berend 2011; Constant, Sigogne, and Watrin 2012; Schneider et al. 2014a). Vincze, Nagy, and Berend (2011) and Constant, Sigogne, and Watrin (2012) handle contiguous MWEs using conditional random fields. Diab and Bhutada (2009) use sequential taggers based on support vector machines to identify idiomatic verb-noun constructions. In addition to traditional features like left and right words, lemmas, and character *n*-grams, they also use multiword named entity placeholders. Schneider et al. (2014a) introduce a slightly more complex tagging scheme, allowing gaps and one-level nesting as shown in Figure 4. They use a linear model trained with the perceptron algorithm.

External MWE resources can have a great impact and can be used in two different ways: (a) they are projected on a corpus in order to build an annotated data set for training (Vincze, Nagy, and Berend 2011); or (b) they are used as a source of features of the statistical model (Constant, Sigogne, and Watrin 2012; Schneider et al. 2014a). For

instance, in order to tackle the lexical sparsity of MWEs, some studies showed interest in integrating lexicon-based features in sequence tagging models. Constant, Sigogne, and Watrin (2012) developed a generic approach to compute features from contiguous MWE lexicons. They use this approach to identify French compounds using conditional random fields, showing significant gains compared with settings without lexicon-based features. This method has also been successfully applied and updated for comprehensive MWE identification in English by Schneider et al. (2014a), who performed fine-grained feature-engineering, designing specific features for different MWE lexicons.

3.3 Evaluation of MWE Identification

The evaluation of MWE identification must take all the challenges mentioned in Section 3.1 into account. Automatic evaluation is possible if we compare the output of the MWE tagger with manually annotated sentences on a given test set using traditional measures such as precision, recall, and F-measure. However, these measures are based on full-MWE comparison and ignore partial matches (Constant, Sigogne, and Watrin 2012). Link-based precision and recall, inspired by the Message Understanding Conference criterion for coreference resolution, can be used to distinguish taggers that can identify only part of an MWE from systems that do not recognize it at all (Schneider et al. 2014a). It is also possible to consider partial matches on MWEs by taking the maximum precision and recall among all possible token alignments between the prediction and the gold standard (Savary et al. 2017).[20]

It is worth noting that current partial matching evaluation metrics are not completely satisfactory and should be further investigated. Indeed, these metrics do not take into account the "importance" of tokens within the whole expression. For instance, if the tagger identifies *into account* instead of the whole expression *take into account*, which is partially correct, it misses the syntactic head of the expression and this will cause a downstream semantic parser to fail. We could therefore assume that partial matching evaluation metrics should strongly penalize the system in that case. In the other extreme case, sometimes missing some tokens or identifying additional tokens might not harm semantic analysis. For instance, in the sentence *John had a bath*, the inclusion of the determiner *a* as a lexicalized element of the light-verb construction *have_bath* is questionable and a tagging error on this element should not be strongly penalized. The other way around, missing *had* and *bath*, which are obligatory elements, should be strongly penalized. Evaluation metrics using weighting schemes that assess partial matches by the "importance" of the tokens should be developed in the future.

Many advances in the development of better evaluation metrics are a by-product of shared tasks on MWE identification. The DiMSUM shared task for joint identification and supersense tagging of MWEs in English has put identification in focus (Schneider et al. 2016). Among the proposed systems, sequence taggers and the use of clustering methods for generalization seem to bring good results. The results of the PARSEME shared task on verbal MWE identification (Savary et al. 2017) confirm that sequence taggers can perform as well as parsing-based approaches, depending on the language and on the proportion of discontiguous MWEs.

20 The main difference between the measures proposed by Schneider et al. (2014a) and Savary et al. (2017) is that the latter was designed to allow single-token MWEs such as compounds (*snowman*) and contractions (*suicidarse* = *suicidar+se*, 'to suicide' in Spanish).

Extrinsic evaluation of MWE identification has also been performed. Identifying (discontiguous) MWEs influences the performance of information retrieval (Doucet and Ahonen-Myka 2004) and word sense disambiguation (Finlayson and Kulkarni 2011). Evaluation of identification when performed before parsing (Section 4.2) or before MT (Section 5.2) is detailed in the corresponding sections.

3.4 Open Issues in MWE Identification

In spite of receiving less attention than MWE discovery, MWE identification has made significant progress. On the one hand, experiments have shown that it is possible to identify specific MWE categories, especially ambiguous ones such as verbal idioms, using mostly unsupervised models (Katz and Giesbrecht 2006; Boukobza and Rappoport 2009; Fazly, Cook, and Stevenson 2009). On the other hand, supervised taggers have been successfully used to learn more general MWE identification models capable of handling several MWE categories simultaneously (Vincze, Nagy, and Berend 2011; Constant, Sigogne, and Watrin 2012; Schneider et al. 2014a).

Achieving broad-coverage MWE identification is still an open issue for both unsupervised and supervised methods. Unsupervised methods are usually evaluated on small data sets and it is unclear to what extent the proposed models are generalizable. Supervised methods require sufficient training data and do not perform well on rare MWEs, which have not been seen often enough in the training data. Integrating the approaches presented in this section, for example, using unsupervised features in supervised taggers, could be a promising research direction to address this issue.

Moreover, current identification models cannot always properly model and recognize discontiguous and overlapping expressions. As for discontiguous MWEs, the use of parsers can help (Section 4). As for overlap, some approaches can deal with nesting (Schneider et al. 2014a) but other types of overlap are considered sufficiently rare to be safely ignored. For example, partial overlapping like in $pay_1\ close_2\ attention_{1,2}$ containing the expressions *pay attention* and *close attention* is usually ignored. Although it is not straightforward to model overlapping MWEs within taggers and parsers, it would be interesting to develop new identification models that can elegantly handle overlap.

The success of MWE identification for languages like English and French has relied heavily on high-quality lexicons and annotated corpora, which are rare resources. Broad-coverage hand-crafted MWE lexicons take years to build, and the use of faster, automatic discovery methods directly for identification, bypassing lexicographers, has not been sufficiently studied. Furthermore, annotated corpora containing MWEs are often constructed for other purposes (e.g., treebanks), and MWE annotation is not always consistent (Green et al. 2011). Even when the annotation was performed explicitly for identification purposes, consistency problems always occur because of the complex nature of MWEs (Hollenstein, Schneider, and Webber 2016). Hence, the development of robust annotation guidelines and coherently annotated corpora is a bottleneck that requires attention in the near future.

The use of end-to-end sequence taggers based on recurrent and/or deep neural networks looks promising (Legrand and Collobert 2016) and remains to be explored. One of the potential advantages of these methods is that they can deal with word vectors (embeddings) that are of a semantic nature. Because MWEs are closely related to semantic compositionality, such models could learn how to tag MWEs when the compositionality of word vectors is breached.

4. MWE Processing and Parsing

Parsing is a historical field of NLP and continues to receive much attention even after decades of research and development. A parser takes as input a sequence of tokens and generally computes one or more grammatical representations. Generally, these take the form of a tree whose structure depends on the grammatical framework (e.g., constituency or dependency trees). Parsers fall into two main classes: grammar-based parsers and grammarless parsers.

Grammar-based parsers rely on computational grammars, that is, collections of rules that describe the language, expressed in a given grammatical formalism like tree adjoining grammar (TAG) (Joshi, Levy, and Takahashi 1975), combinatory categorial grammar (Steedman 1987), lexical functional grammar (LFG) (Kaplan 1989), and head-driven phrase-structure grammar (HPSG) (Pollard and Sag 1994). Grammars may also be composed of sets of finite-state rules that are incrementally applied (Joshi and Hopeli 1996; Ait-Mokhtar, Chanod, and Roux 2002). Two different strategies are generally used to handle ambiguity. In the first strategy, the process comprises two phases: an analysis phase, generating a set of possible syntactic trees, followed by a disambiguation phase based on heuristics (Boullier and Sagot 2005; Wehrli 2014) or statistical models (Riezler et al. 2002; Villemonte De La Clergerie 2013). In the second (mainstream) strategy, the grammar is accompanied by a statistical model. For instance, parsers based on generative-models assign probabilities to rules of an underlying grammatical formalism, as in probabilistic context-free grammars (PCFGs) (Charniak and Johnson 2005), tree-substitution grammars (Green et al. 2011), TAG (Resnik 1992), and LFG (Cahill 2004). The parsing algorithms generally rely on dynamic programming. They usually include one pass, but two-pass processes also exist. For instance, Charniak and Johnson (2005) successfully propose applying a discriminative reranker to the n-best parses produced by a generative PCFG-based parser.

Grammarless parsing is performed without any underlying grammar and is based on discriminative approaches. It uses machine learning techniques only, mainly (not exclusively) in the dependency framework. The different parsing algorithms vary from local search approaches, such as transition-based systems (Nivre, Hall, and Nilsson 2004), to global ones, such as graph-based systems (McDonald et al. 2005).

For both main classes of parsers, significant progress has been made using deep learning techniques (Chen and Manning 2014; Durrett and Klein 2015; Dyer et al. 2015; Pei, Ge, and Chang 2015).

MWE-aware parsing comprises a very tiny portion of this abundant literature. Most MWE-aware parsing strategies are adaptations of standard parsers that experiment with various orchestration schemes for identification. Depending on the scheme, adaptations include modifications to training data, grammatical formalisms, statistical models, and parsing algorithms, as well as specialized modes of interaction with lexicons.

4.1 Motivations and Challenges for MWE-Aware Parsing

The motivation for MWE-aware parsing is 3-fold: (1) to improve the syntactic parsing performances on sentences containing MWEs (both on internal MWE structure and on the surrounding sentence structure), (2) to improve MWE identification performance, and (3) to improve MWE discovery performance. The latter two items rely on the fact that processing of some MWEs hinges on their syntactic analysis. Parsing faces different challenges with respect to identification because of non-compositionality and ambiguity (Section 1.2). Paradoxically, both challenges may be tackled using parsing. Besides

these, MWE-aware parsing addresses the other MWE-related challenges discussed in Section 1.2.

Ambiguity. An MWE might be ambiguous between accidental co-occurrence and literal and idiomatic uses. An incorrect identification can mislead the parser. In particular, complex function words that have a key role in syntax may be ambiguous (e.g., *up to*). For instance, in *John looked up to the sky*, the sequence *up to* should not be identified as a multiword preposition. If so, it would prevent the right analysis: *(John) ((looked) (up to the sky))* instead of *(John) ((looked) (up) (to the sky))*.

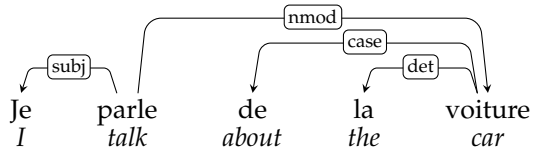
Conversely, combining MWE identification and parsing can help resolve such ambiguities, yielding both better identification and parsing models. Multiword function words such as complex prepositions, conjunctions, and adverbials (*up to*, *now that*, *by the way*) can be disambiguated by their syntactic context (Nasr et al. 2015). For example, the sequence *de la* in French can be either a compositional sequence (preposition *de* + determiner *la*), or a complex partitive determiner, as shown in the following examples and their corresponding syntactic analyses in Figure 5:

(1) Je parle de la voiture
I talk about the car

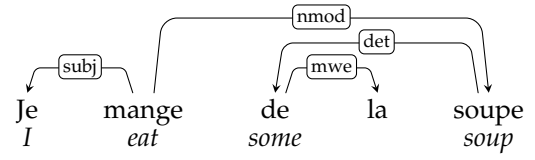
(2) Je mange de la soupe
I eat some soup

MWE-aware parsing is a natural way to solve this ambiguity. The intransitive verb *parle* (*talk*) selects the preposition *de* (*about*), whereas *mange* (*eat*) requires a noun phrase as its object. Furthermore, one of the main challenges of parsing in general is attachment ambiguity. As MWEs tend to form full syntactic constituents, their identification can guide attachment decisions (Wehrli, Seretan, and Nerima 2010).

Non-compositionality. MWEs can be defined as exceptions to regular composition rules. This non-compositionality can take place at different levels: morphological, distributional, syntactic, semantic, and pragmatic. In particular, some expressions display syntactic irregularity in their internal structure—that is, they are irregular with respect to a grammar. Therefore, if the parser is not aware of the existence of such cases, the analysis will fail. For instance, the adverbial *by and large* is the coordination of a



(a) Compositional analysis



(b) Complex partitive determiner analysis

Figure 5
Syntactic analyses of the French sequence *de la* using the universal dependencies annotation scheme.

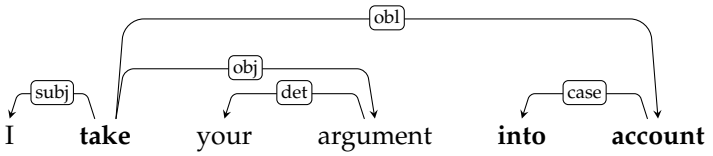


Figure 6
Syntactic analysis of the discontiguous verbal MWE *take into account* using the universal dependencies annotation scheme.

preposition and an adjective that is an irregular pattern for a syntactic constituent in English.

Conversely, depending on the type of parser used, some types of syntactic non-compositionality can be captured by a parser as, like *by and large*, they violate standard grammatical rules. For example, informing a discriminative parser with POS *n*-gram features may help capture such non-compositionality. Here too, higher levels of non-compositionality are not directly relevant for parsing because they do not interfere with the resulting parse tree, even though information about non-compositional MWEs can guide the parser.

Discontiguity. Some MWEs can appear in discontiguous configurations, like *give up* in *John gave it up*, and *take into account* in *Mary took your argument into account*. The identification of such discontiguous MWEs can hardly be handled by purely sequential approaches (Section 3.2.3), except maybe for special cases when gaps are short (Schneider et al. 2014a). Because syntactic links can relate non-adjacent words, parsing may help the identification of discontiguous MWEs (Seretan 2011). In Figure 6, representing the syntactic analysis of a discontiguous instance of the expression *take into account*, the verb *take* is a syntactic neighbor of the noun *account*, which should facilitate the identification of the whole expression.

Variability. Flexible MWEs may undergo syntactic variations. For instance, light-verb constructions accept verbal inflection (*make/made/making a decision*), passivization (*John made a decision* $\rightarrow$ *a decision was made by John*), and insertion of free modifiers (*John made an important decision*). As parsers provide syntactic structure, they can be useful to capture and aggregate MWE variants for discovery (Seretan 2011). In MWE identification, though, parsers need to take variability into account when matching MWE dictionary entries with their instances in text.

The main challenge for MWE-aware parsing is the orchestration of MWE identification and parsing, that is, the question of *when* one task should be performed with respect to the other (Section 4.2).

4.2 Orchestration of MWE Processing and Parsing

This section describes the different orchestration strategies involving syntactic parsing and MWE processing under four subsections as a consequence of the interactions between parsing, identification, and discovery shown in Figure 1: discovery after parsing, and identification before, during, and after parsing. In particular, it shows how the different challenges (Section 3.1 and Section 4.1) can be handled, and summarizes the reported advantages and disadvantages of each strategy.

4.2.1 Discovery after Parsing. Discovery can be fruitfully informed by syntactic structure and can help tackle challenges like discontiguity and variability (Section 2.1), as MWE

components usually belong to the same syntactic constituents. Therefore, having the syntactic structure may help capture MWEs with non-adjacent elements. For instance, Seretan (2011) empirically shows that such information enables the extraction of more relevant MWE candidates as compared to standard extraction methods based on fixed-size windows. Moreover, discovery of flexible constructions usually requires linguistic patterns based on parsing: For instance, verbal expressions combining a head verb and a noun or a prepositional phrase (Fazly and Stevenson 2006; Cook, Fazly, and Stevenson 2007; McCarthy, Venkatapathy, and Joshi 2007; Krenn 2008; Seretan 2011), verb-particle constructions (Bannard 2002; McCarthy, Keller, and Carroll 2003), or constructions formed from a head noun and a prepositional phrase (Weller and Heid 2010). For an interesting illustration, Baldwin (2005) proposes different morphosyntactic and syntactic patterns to extract English verb-particle constructions with valence information from raw text. In particular, he shows the effect of using the outputs of a POS tagger, a chunker, a chunk grammar, or a parser, either individually or combined via a classifier. It experimentally appears that the ensemble method significantly outperforms the individual performances. As for individual scores, the use of shallow syntactic information like chunks tends to be prevalent.

It is also possible to use pattern-free approaches like Martens and Vandeghinste (2010) and Sangati and van Cranenburgh (2015), who propose discovery methods not dedicated to a specific MWE category but based on recurring tree fragments and association measures.

4.2.2 Identification Before Parsing. When MWE identification is performed before parsing, the search space of the parsing algorithm is reduced. Hence, the main advantage of this type of orchestration is that the parsing process becomes less complex. The parser takes as input a sequence of partially analyzed linguistic units. This can be seen as a **retokenization** process, where the pre-identified MWE is merged into a single token (e.g., *by the way* → *by_the_way*). MWE identification prior to parsing has been implemented both in statistical (Cafferkey, Hogan, and van Genabith 2007; Korkontzelos and Manandhar 2010; Constant, Sigogne, and Watrin 2012; de Lhoneux 2015) and rule-based parsers (Brun 1998; Mamede et al. 2012).

This orchestration type has the advantage of simplicity and empirical efficiency. For instance, Cafferkey, Hogan, and van Genabith (2007) show that pre-identifying multiword named entities and prepropositional MWEs improves parsing accuracy in the constituent framework. The best system of the track on MWE-aware dependency parsing in the SPMRL 2013 shared task (Seddah et al. 2013) was the only one that included deterministic pre-identification (Constant, Candito, and Seddah 2013).

Limitations. The pre-identification approach suffers from limitations. First, in this scenario, most of the proposed methods are limited to contiguous MWEs. Handling discontiguous ones may involve word reordering: *John **gave** it **up*** → *John **gave_up** it*. In addition, when MWE components are concatenated into a single token, their internal syntactic structure is lost, whereas it may be required for the semantic processing of semi-compositional MWEs. However, this can be performed *a posteriori*, for instance, by applying simple rules based on POS patterns (Candito and Constant 2014).

Then, the retokenization increases data sparsity that negatively affects parsing performance, because the vocabulary size increases whereas the total amount of training data is the same. Eryiğit, İlbay, and Can (2011) showed that the concatenation operation of different MWE categories has different impacts on parsing performance for Turkish. Whereas retokenization of multiword named entities and numerical expressions

improved dependency parsing performance, retokenization of light-verb constructions harmed it.

Another disadvantage is that pre-identification is deterministic, so the syntactic parser cannot recover from MWE identification errors. A sentence like *He walked by and large tractors passed him* cannot be analyzed correctly if *by and large* is pre-analyzed as a multiword adverb (*by_and_large*). Errors arise mainly due to challenging aspects like ambiguity (accidental co-occurrence identified as an MWE) and variability (an MWE missed because it is different from the canonical form in the lexicon, see Section 3.1).

Non-Deterministic Approaches. To fully profit from using MWE identification as a preprocessing step, parsing has to be carried out non-deterministically, since several alternatives should be maintained and eventually resolved using a disambiguation model of some kind. Rule-based parsers deal with this problem to some extent by taking as input a lattice of possible POS sequences and MWE segmentations constructed from a lexicon-based preprocessing phase (Villemonte De La Clergerie 2013; Sagot and Boullier 2005). In the statistical paradigm, Constant, Le Roux, and Sigogne (2013) successfully used an MWE tagger based on conditional random fields that generates the most probable outputs in the form of a lattice of lexical units. The lattice is then fed into the parser, which is in charge of selecting the best lexical segmentation, as well as the best syntactic tree. With this rationale, the MWE identification module of Urieli (2013) successfully feeds the initial beam of a transition-based parser with its n-best segmentations, each associated with a score. These proposals suggest that a crucial aspect of MWE identification in MWE-aware parsing systems is whether ambiguous analyses can be handled.

4.2.3 Identification During Parsing. As mentioned earlier (Sections 3.1 and 4.1), MWEs may be discontiguous and their identification may require access to the syntactic environment available during parsing. Some studies have shown the great interest of performing MWE identification during syntactic parsing. We henceforth also use the term "joint" to refer to such approaches.

Joint Grammar-Based Approaches. In a grammar-based parser, MWE identification is often integrated in the grammar. MWEs are generally found in a lexical resource, and parsers embody mechanisms to link MWE entries to grammar rules, as in Abeillé (1995) for TAG, Attia (2006) for LFG, and Copestake et al. (2002) and Villavicencio et al. (2007) for HPSG. For instance, in the TAG paradigm, Abeillé and Schabes (1989) link MWE lexical entries to tree rules that are anchored by multiple components present in the lexical entry. This mechanism has been integrated in the XTAG project that aims to construct a lexicalized TAG for English (XTAG 2001). In practice, rule-based parsers can also use MWE identification as a cue to locally select the best syntactic analysis: for instance, Wehrli (2014) applies heuristics favoring MWE analyses.

Where statistical grammar-based parsers are trained from a reference treebank, MWEs must be annotated within the treebank. Typically, each MWE is annotated with a specific subtree having a flat structure (Arun and Keller 2005; Green et al. 2011; Seddah et al. 2013). In particular, Green et al. (2011) and Green, de Marneffe, and Manning (2013) learned PCFG and probabilistic tree-substitution grammars from such treebanks. Sangati and van Cranenburgh (2015) successfully learned double data-oriented parsing models for which extracted subtrees only count if they happen to form a largest shared fragment from another pair of trees in the treebank.

Joint Grammarless Approaches. In the case of grammarless parsers, seminal studies have confirmed the promise of joint statistical models for handling the parsing and identification tasks at the same time (Nivre and Nilsson 2004; Eryiğit, İlbay, and Can

2011). In the dependency parsing framework, the model is trained on a treebank where MWEs are annotated in the form of a subtree and specific arcs denote MWE components with either a shallow structure (Nivre and Nilsson 2004; Eryiğit, İlbay, and Can 2011; Seddah et al. 2013; Kong et al. 2014; Nasr et al. 2015), or a deeper one (Vincze, Zsibrita, and Nagy 2013; Candito and Constant 2014). For instance, Vincze, Zsibrita, and Nagy (2013) incorporate syntactic functions in the light-verb construction labels in order to retain the internal structure of the expression.

Such approaches generally use off-the-shelf parsers that perform well for MWE-aware parsing as they are informed by many lexicalized features.[21] Such workaround approaches operating on very simple representations of MWEs have some limitations. Syntactic analysis and MWE identification are handled using the same types of mechanisms, whereas their underlying linguistic models are quite different, though interleaved. Some proposals try to handle this problem. For instance, in the dependency framework, Nivre (2014) mentioned the possible integration of a new action, namely, a transition dedicated to MWE identification, within a transition-based parsing system. In this line of research, Constant and Nivre (2016) proposed a transition-based system for joint lexical and syntactic analysis including specific transitions for MWE identification using a dual representation of MWEs (syntax-irregular vs. syntax-regular MWEs).

On the Use of Lexicons. The MWE identification section (Section 3) has shown that the use of lexicons increases MWE identification performances. Joint rule-based grammar-based MWE-aware parsing generally embodies mechanisms to link the grammar to a lexicon, as illustrated above for the TAG formalism. The use of lexicons is more complicated for grammar-based parsers based on generative statistical models. That is why their integration within non-deterministic processing chains is achieved by including support from MWE taggers that use lexicons (Sections 4.2.2 and 4.2.4).

One great advantage of joint grammarless MWE-aware parsers based on discriminative models is that they can be informed by external lexicons using additional features. For instance, Candito and Constant (2014) showed that incorporating MWE features based on MWE lexicons greatly improves the accuracy. Nasr et al. (2015) also showed that incorporating specific syntactic subcategorization lexicons helped the disambiguation of ambiguous complex function words. For instance, in French, the sequence *bien que* is either a multiword conjunction (*although*) or the literal sequence composed of an adverb (*well*) followed by a relative conjunction (*that*). This ambiguity may be resolved using the verb in the syntactic neighborhood. The authors included specific features indicating whether a given verb accepts a given grammatical element: *manger (to eat)* -QUE -DE, *penser (to think)* +QUE -DE, *boire (to drink)* -QUE -DE, *parler (to speak)* -QUE +DE. The QUE feature indicates whether the verb accepts a clausal complement introduced by *que* (that), and the DE feature indicates whether the verb accepts a nominal complement introduced by preposition *de* (of, from).

Discussion. Joint approaches are of great interest for MWEs having syntactic variability. In particular, Eryiğit, İlbay, and Can (2011) and Vincze, Zsibrita, and Nagy (2013) state that a joint approach using a dependency parser is very successful for the identification of light-verb constructions in Turkish and in Hungarian, respectively. Nonetheless, such approaches have the inconvenience of complicating the parsing stage through an increase in the size of the label sets. For instance, the literature shows

21 Also of note is the work of Finkel and Manning (2009), which is limited to named entity recognition and constituent parsing: They jointly performed both tasks using a parser based on conditional random fields, combining features specific to both tasks. The experimental results showed that the accuracy of both tasks increased.

mixed results for non-compositional open compounds for which a pre-identification approach is sometimes more accurate than a joint one (Constant and Nivre 2016). The right balance therefore has to be found.

An interesting way to deal with this issue is to combine the *before* and *during* strategies using dual decomposition, as shown to be very effective in parsing tasks (Martins et al. 2010; Le Roux, Rozenknop, and Foster 2013).[22] In the case of MWE-aware parsing, results can be improved using dual decomposition combining MWE pre-identification with a joint parser. For instance, Le Roux, Rozenknop, and Constant (2014) combine several sequential MWE taggers based on conditional random fields with several constituent MWE-aware joint parsers. All taggers and parsers are trained independently, and the system iteratively penalizes each parser and tagger until agreement on MWE segmentation is reached. Such an approach reaches state-of-the-art results for French MWE identification and parsing. Nonetheless, one drawback of their approach is that it only handles contiguous MWEs. It is interesting to note that the dual decomposition approach makes it possible to use several shallow MWE annotation schemes. For instance, taggers use IOB annotations, and parsers use constituent subtree annotations. From both schemes, one can compute the predicted MWE spans used for computing the agreement between MWE-aware systems. In light of inspiring results regarding shallow MWE annotations, we conclude that the dual decomposition approach may also benefit from other shallow MWE annotations. This hypothesis deserves future investigation.

4.2.4 Identification after Parsing. Although it is well known that predicting syntactic structure might help tackle challenges like discontiguity and variability, especially for verbal expressions, very few studies have experimented with identification after parsing. Fazly, Cook, and Stevenson (2009) used a parsed text in order to identify verb-noun idiomatic combinations. Nagy and Vincze (2014) also successfully use such an approach to identify verb-particle constructions in English. They positioned a classifier on top of a standard parser in order to select verb-particle constructions from a set of candidates extracted from the parsed text. Baptista et al. (2015) identify verbal idioms in Portuguese using the incremental finite-state parser XIP (Ait-Mokhtar, Chanod, and Roux 2002) as part of the STRING NLP pipeline (Mamede et al. 2012). The finite state parser first recognizes chunks then identifies syntactic relations between them by incrementally applying hand-crafted rules. Then, new rules are added in order to capture verbal idioms based on already predicted lexical and syntactic information. Also, most of the systems proposed in the PARSEME Shared Task (Savary et al. 2017) used tagging supervised models relying on syntactic features in order to identify verbal MWEs.

Instead, most identification methods are based on less complex preprocessing stages (tokenization, POS tagging, lemmatization, etc.), as shown in Section 3. One reason could be that discontiguous MWEs tend to have small gaps, as shown in English by Schneider et al. (2014a). Another reason could be that parsers are error-prone and error-propagation might harm MWE identification (unlike the case for discovery where errors are compensated by large quantities of data). To tackle the problem of errors, an interesting approach is to make use of reranking (Charniak and Johnson 2005). For instance, Constant, Sigogne, and Watrin (2012) used an MWE-dedicated reranker on top

22 Dual decomposition is a combinatorial optimization approach that consists in dividing a problem into subproblems having agreement constraints. In particular, it enables one to efficiently combine several systems having a common subtask.

of a parser generating the *n*-best parses (including MWE identification) and showed significant improvement in MWE identification accuracy.

4.3 Evaluation of MWE-Aware Parsing

Evaluating a syntactic parser generally consists of comparing the output to reference (gold standard) parses from a manually labeled treebank. In the case of constituency parsing, a constituent is treated as correct if there exists a constituent in the gold standard parse with the same labels and starting and ending points. These parsers are traditionally evaluated through precision, recall, and F-measure metrics (Black et al. 1991; Sekine and Collins 1997).

In standard dependency parsing with the single-head constraint,[23] the number of dependencies produced by the parser should be equal to the number of total dependencies in the gold-standard parse tree. Common metrics to evaluate these parsers include the percentage of tokens with correct head, called **unlabeled attachment score** (UAS), and the percentage of tokens with correct head and dependency label, called **labeled attachment score** (LAS) (Buchholz and Marsi 2006; Nilsson, Riedel, and Yuret 2007).

The evaluation of identification and discovery has been discussed in previous sections. However, evaluation of MWE-aware parsers and of whether or not MWE identification helps to improve the parsing quality requires some additional care. In most work where MWE identification is realized before parsing, the MWEs are merged into single tokens (Section 4.2.2). As a result, the common metrics for parsing evaluation given above become problematic for measuring the impact of MWE identification on parsing performance (Eryiğit, İlbay, and Can 2011). For example, in dependency parsing, the concatenation of MWEs into single units decrements the total number of evaluated dependencies. It is thus possible to obtain different scores without actually changing the quality of the parser, but simply the representation of the results. Instead of UAS and LAS metrics, the attachment scores on the surrounding structures, namely, UAS_{surr} and LAS_{surr} (i.e., the accuracies on the dependency relations excluding the ones between MWE elements) are more appropriate for the extrinsic evaluation of the impact of MWE identification on parsing. Similar considerations apply to constituency parsing.

Although UAS_{surr} and LAS_{surr} are valuable metrics for measuring the impact of different MWE categories on parsing, they are troublesome with automatic MWE identification when gold-standard MWE segmentation is not available, because erroneously identified MWEs would degrade parsing scores on the surrounding dependencies.

An alternative solution is to detach the concatenated MWE components (if any) into a dependency or constituency subtree (Candito and Constant 2014; Eryiğit, İlbay, and Can 2011). In this way, the standard evaluation metrics are still applicable in all different orchestration scenarios and work on both contiguous and non-contiguous cases, thus providing a means to assess the performance of joint syntactic parsing and MWE identification as a whole.

4.4 Open Issues in MWE-Aware Parsing

A few studies have established that optimal parsing of different MWE categories requires different treatment and different orchestration scenarios (Section 4.2). In order to

23 Each dependent node has at most one head in the produced dependency tree.

design a method for finding the optimal orchestration, a more systematic investigation needs to be carried out involving the three MWE identification positions for every MWE category in every language.

To avoid the data sparsity problem caused by the concatenation strategy used in the MWE pre-identification scenario, another strategy that deserves investigation is constrained parsing (Nivre, Goldberg, and McDonald 2014). In these systems, the constraints are incorporated into the parsing system as a set of preconditions that force the parser to keep the given dependencies/constituents in the output parse.

Studies on statistical MWE-aware parsing tend to work on very simple representations of MWEs. The benefit of adopting more complex, deeper representations capable of representing, for example, embedded MWEs (Finkel and Manning 2009; Constant, Le Roux, and Tomeh 2016; Constant and Nivre 2016), is as yet unclear. There is a case to be made for such approaches to be investigated more deeply on data sets with comprehensive MWE annotations in many different languages.

Previous sections have shown the different ways identification and parsing can interact. In particular, they show the great interest of using manually validated MWE lexicons. We could also imagine how MWE lexicons extracted from discovery might be directly integrated in the MWE-aware parser. Such methods are very close to those that integrate lexical affinities, acquired from large quantities of external raw data, into a standard statistical parser (Volk 2001; Bansal and Klein 2011; Mirroshandel, Nasr, and Le Roux 2012; Schneider 2012).

5. MWE Processing and Machine Translation

MT systems aim to automatically translate a source text into a target text that retains the meaning and fluency of the source. They must take into account the lexical, morpho-syntactic, syntactic, and semantic constraints in source and target languages. The main MT paradigms are summarized here.

Statistical machine translation (SMT) acquires the ability to translate from parallel data using machine learning techniques. Like all such systems, it includes a *training phase*, which uses the data to build probabilistic models, and a *decoding phase*, where these models are deployed to actually carry out translation of an unseen source language sentence.

During training, two kinds of probabilistic model are built: a *translation model*, derived from bilingual corpora, and a *language model*, from monolingual corpora. Both models assign probabilities: the translation model to source/target language fragments, and the language model to target language word sequences (Koehn 2010).

During decoding, the system generates many hypothetical translations for each source sentence and chooses the most probable hypothesis. This is calculated for each by combining probabilities assigned by the acquired translation and target language models. The effective performance of this calculation requires considerable ingenuity, given the exponential number of possible translations and orderings of translated sentence fragments, the non-trivial computations involved, and the real time and memory constraints.

Variants of these steps take into account contiguous sequences of words in so-called phrase-based SMT (Koehn, Och, and Marcu 2003), syntactic structures in syntax-based SMT (Chiang 2007; Hoang and Koehn 2010), or linguistic annotation layers in factor-based SMT (Koehn and Hoang 2007)).

Phrase-based SMT and its variants build *phrase tables*—that is, a list of source fragments (words, phrases, subtrees), their translations, and their translation probabilities

that take into account word sequences, not only simple words. In principle, therefore, such systems can naturally handle contiguous MWEs. Whether they can handle them correctly in all cases is, of course, a separate question.

More recently, neural machine translation (Kalchbrenner and Blunsom 2013; Cho et al. 2014) proposes alternative methods to compute translation probabilities, by using recurrent neural networks to model the translation task. Most neural translation systems use an encoder–decoder architecture. The input sentence is encoded into a fixed-length or variable–length vector and then one or more decoders use this representation to obtain the target sentence. The probability of the translation of one word is computed on the basis of the translation probabilities of previous words. An attention model is frequently used to represent larger contexts for the translated words and sentences. Indeed, attention models represent source word and larger-context words (using a dot product of vectors or multilayer perceptrons) to generate a target word. Few neural machine translation systems take into account fine linguistic descriptions (Sennrich and Haddow 2016). Neural machine translation obtains impressive improvements of the evaluation scores such as BLEU (Wu et al. 2016).

Rule-based machine translation (RBMT) uses large lexicons and explicit rules describing the syntactic and semantic constraints on both the source and the target language. Transfer rules are used to map source language structures to target language ones and to identify the right translation. These rules are based on formal grammars or intermediate language-independent structures (such as minimal recursion semantics [Oepen et al. 2004]) capable of generating correct translation equivalents.

Finally, **example-based machine translation (EBMT)** is based mainly on examples in the form of large translation memories (large collections of source/target sentence pairs) but also uses rules to acquire new linguistic knowledge dynamically. EBMT is based on a translation by analogy approach, where at run time translations are obtained by looking up and using examples stored in translation memories. The translation process is organized in three stages: (i) matching of input sentences with translations previously stored, (ii) retrieval of these translations, and finally (iii) adaptation or recombination of the target sentences. An early review of EBMT appears in Somers (1999).

In this introductory section we defined the various approaches used in machine translation and their acronyms. In the following sections we will use the acronyms SMT, RBMT, and EBMT instead of the complete terms to improve text readability.

5.1 Motivations and Challenges for MWE-Aware Machine Translation

The main motivation for MWE-aware MT is that none of these paradigms can consistently address the basic features of MWEs, some of which were already mentioned in Section 1.2: ambiguity, discontiguity, non-compositionality, and variability. We briefly review these challenges here.

Ambiguity. Here the components of an MWE can be interpreted and translated literally or idiomatically according to the context, and the two readings are easily confused. For example, the French MWE *jeter l'éponge* (lit. *to throw the sponge*) means *to resign*. However, the expression might be used literally (the person is washing the dishes and literally throws the sponge). The challenge concerns the nature of the information required to make the right choice, and how to represent it. In parsing we should note that this is likely to include more contextual, extra-linguistic, or multilingual information than is available in most MT systems.

Discontiguity. Translation is hampered by alien elements occurring between the components of an MWE. Google Translate renders *John picked up the book* as *John ramassa*

le livre in French, which is correct, if a little literary. But *John picked the book up* is translated as *John prit le livre jusqu'à*, which is ungrammatical. The challenge, again, concerns the information required for such discontiguous MWEs to be recognized, and how that information is brought to bear. A pertinent issue is whether there are special cases to be exploited, as is the case for phrases having a fixed or semi-fixed frame with slots for one or more fillers, such as *give [. . .] a [. . .] break* or *on the one hand, [. . .] on the other hand*.

Non-compositionality. This implies that a compositional translation strategy will generally fail. Besides resolving the ambiguity of whether a given case is compositional there is another challenge that arises because compositionality is not necessarily all or nothing: It has degrees. At one extreme, we have non-compositional expressions like *red tape*, meaning excessive bureaucracy. At the other we have compound nouns like *honeymoon* for which a compositional translation may be correct (e.g., *luna di miele* in Italian). In between the two, we have semi-compositional expressions such as *to take a decision*. The challenge is whether the degree can be predicted and exploited, because if an MWE is *mostly* compositional, then a mostly compositional strategy stands a chance of producing an acceptable translation. There are special cases where semi-compositional translation strategies work. So *ask a question* should be translated by *a pune o întrebare* (lit. *put a question*) in Romanian and not by **a cere o întrebare*. The verb *to ask* could not be translated in Romanian as *a cere/to demand* but as *a pune/to put* due to the specific lexical constraints: *a cere/to demand* could not select the noun *întrebare/question*, so the other synonym is selected. The challenge is to select the right translation by taking into account all the constraints.

Variability. MT systems must identify and translate all variants of an MWE. Some variants can be challenging to recognize, particularly when they involve syntactic/semantic constraints. For example, the expression *he has cooked his goose* means that he has fallen into a trap of his own making. But for this reading, *he* and *his* must corefer. If they do not, the reading is compositional. Not only is this condition tricky to verify, but there are variants—for example, involving number and gender: *she has cooked her goose*—whose identification might require hand-made or automatically learned rules. Once these are in place, variants of MWEs can be identified using methods presented in Section 3.2.

Translation asymmetries. These occur when an MWE in the source language is not necessarily translated by an MWE in the target language and vice versa. They represent an additional challenge. In general, we can have different correspondences, exemplified by the following English–Italian examples: many-to-many (*to kick the bucket* → *tirare le cuoia*), many-to-one (*to kick the bucket* → *morire*), and one-to-many (*svegliare* → *to wake up*). There are several challenges. One is to decide whether to choose an MWE target when a one-word alternative exists. Another is to determine what syntactic adjustments must be made to the target to retain fluency with asymmetric output. Another challenge is how best to exploit asymmetry for the purpose of discovery, as discussed further in Section 5.2.1.

Some of these challenges might be indirectly handled by specific types of MT systems with various orchestration strategies. Idiomatic expressions or contiguous MWEs might be correctly translated in phrase-based SMT (Cap et al. 2015) or by neural MT (Wu et al. 2016), which take into account larger contexts. However, ambiguous MWEs are often translated in SMT with their literal meaning because of a larger translation probability. Syntax-based SMT might capture frequent terms or light-verb constructions without distinguishing MWEs from simple noun-noun or verb-noun combinations. Neural MT systems handle some specific compounds by specific segmentation strategies, but no discontiguous MWEs.

EBMT systems handle contiguous MWEs through their specific translation strategies. RBMT designs specific hand-made rules to translate MWEs, but ambiguity is still a problem.

MT systems with various degrees of MWE-awareness have striven to address these challenges to improve the quality of translation (Ren et al. 2009; Kordoni and Simova 2012; Ramisch, Besacier, and Kobzar 2013; Barreiro et al. 2014). The results of these strategies, presented in the next section, vary across language pairs or MWE categories. For example, Ren et al. (2009) report small improvements of the BLEU score for domain-specific terminology, Cap et al. (2015) report significant improvements of the BLEU score for German light-verb constructions, and Pal, Naskar, and Bandyopadhyay (2013) found significant BLEU improvements for English and Bengali multiword named entities. This variation can be explained by the inadequacy of measures used to evaluate MT (e.g., BLEU) for checking the quality of MWE translation. Section 5.2 presents these attempts in terms of some specific orchestration strategies linking MWE processing with MT.

5.2 Orchestration of MWE Processing and MT

As with parsing, we can define several orchestration scenarios for MWE processing with respect to translation. Theoretically, the options are MWE discovery or identification before, during, or after MT. However, the literature is not very systematic or complete in its coverage, so it is difficult to emerge with a very clear picture of their relative effectiveness. In the following discussion, we concentrate on three particular orchestration scenarios: discovery after MT, identification before MT, and identification during MT. Where appropriate, these are divided according to particular categories of MWE and for specific MT paradigms.

5.2.1 Discovery after MT. Figure 1 shows an arrow going from MT to discovery, meaning that MT can support discovery. This has a particular interpretation: It is not the output of MT that provides the support for discovery. Instead, it is the underlying resources (e.g., parallel corpora) and algorithms (e.g., word alignment in SMT approaches) for MT that are shared by discovery, and these are produced beforehand. Hence, discovery is performed *after* MT.

MWE discovery methods based on parallel corpora use alignment to verify whether a source MWE translates word-for-word or rather as a unit on the target side. Alignment is used for both **multilingual discovery**, that is, finding new translation correspondences between MWEs in two languages, and **monolingual discovery**, that is, finding new MWEs in one language based on translation asymmetries. In fact, multilingual data could help monolingual discovery for less-resourced languages.

Multilingual discovery. Word-level alignment (Och and Ney 2000) indicating that groups of two or more source words are frequently aligned with a single target word can indicate potential MWE candidates. For instance, the French compound *appareil photo* will be aligned with the English word *camera*. MWE candidates can therefore be extracted by using many-to-one alignments as in Caseli et al. (2010). However, because automatic word alignments tend to be very noisy, several techniques have been proposed to filter them. Caseli et al. (2010) used predefined POS patterns and frequency thresholds to obtain bilingual MWE lists, as presented in Section 2. Bouamor, Semmar, and Zweigenbaum (2012b) use an association measure to find the translations of each MWE in the target language counterpart without exploiting the alignment.

Parsed bilingual data has also been used to filter word-aligned MWE candidates. Thus, Zarrieß and Kuhn (2009) propose a method for detecting verb-object

MWEs in both source and target languages that are dependency-parsed, only retaining MWEs whose words are bilingually aligned and monolingually linked by syntactic dependencies.

Tsvetkov and Wintner (2014) proposed supervised classifiers to distinguish MWEs from non-MWEs, using linguistically motivated features such as *literal translatability* derived from simple word alignments in parallel corpora. Here, a check is carried out to assess whether the MWE candidate could be literally translated from a bilingual dictionary. Similarly, Rondon, Caseli, and Ramisch (2015) model translatability as the probability of translating the words of the MWE from Portuguese into English and then back to the same Portuguese word.

Monolingual discovery. For *monolingual* discovery we consider the possibility of using translation asymmetries identified in parallel corpora to compile lists of potential MWE candidates in one specific language without using precomputed word alignments. For example, Sinha (2009) discover Hindi by compiling a list of Hindi light verbs and then looking at mismatches (indicating the use of a verb in a more idiomatic sense) in meaning in the corresponding English counterpart, given a list of literal translations of Hindi light verbs into English.

This approach has also been extended to comparable corpora (texts from the same domain, genre, or type that are not in a translation relation). Morin and Daille (2010) collect bilingual terminology from comparable corpora with the help of a bilingual dictionary. This method applies a compositional word-for-word translation for an MWE candidate and searches the most probable translation in the comparable corpus, identified by a term extraction method. If the compositional translation strategy fails, derivation methods are used to find the nearest word in the dictionary and to find the potential translation.

One advantage of all these methods is that they use relatively well-behaved technologies that exploit translation asymmetries together with non-literal translatability to propose multilingual pairs of MWE candidates as well as monolingual MWEs. Unfortunately, they tend to require large parallel or comparable corpora to find appropriate candidates, and these resources are generally lacking.

5.2.2 Identification before MT. The training process in a standard phrase-based SMT[24] system consists of two steps: word alignment and phrase-table construction. Here, identification takes the form of a preprocessing step (before the training process) that transforms MWEs into an intermediate representation: single units, where component words are concatenated by an underscore character (Carpuat and Diab 2010),[25] a definition from the dictionary (Salton, Ross, and Kelleher 2014),[26] or a paraphrase (Ullman and Nivre 2014).[27] Such preprocessing methods identify MWEs in the monolingual parts of corpora by using external resources (bilingual dictionaries, term databases, or parallel corpora), by applying MWE identification tools (Section 3), or a combination of both (Ghoneim and Diab 2013).

SMT systems. In SMT, MWE replacement takes place before the word alignment step (the "static" approach according to Carpuat and Diab [2010]), using rule-based

24 Phrase-based SMT are the most popular systems among SMT. The orchestration strategies presented in this section and the following one apply to phrase-based SMT.

25 The Romanian term *aparat de fotografiat* for *camera* becomes *aparat_de_fotografiat*.

26 The French idiom *jeter l'éponge* lit. *to resign* is replaced by its meaning *abandonner*.

27 The Swedish compound for railway station *järnvägsstation* is rephrased as *station för järnväg* (lit. *station for railway*).

MWE identification and lexicon look-up with monolingual general dictionaries, bilingual dictionaries, and specific bilingual lists containing multiword named entities or terms and their translations.

When such resources are not available or incomplete, lists of MWE candidates obtained by MWE discovery methods can be applied directly to transform MWEs into single tokens. This strategy handles not only contiguous MWEs, but also their inflected variants. Some specific cases of discontiguous MWEs such as light-verb constructions might be handled by a specific annotation of the light verb (Cap et al. 2015).

EBMT and RBMT systems. For both EBMT and RBMT, lexical resources are essential for handling contiguous MWEs (Barreiro 2008). EBMT uses translation memories to properly identify these, for which word sequences, sometimes based on parsed data (Kim, Brown, and Carbonell 2010), are listed as possible translations. RBMT also uses compositional rules that are combined with transfer rules to handle syntactic variants and discontiguity for some MWEs (Forcada et al. 2011). In the compositional approach, MWE handling is obtained by means of tagging and syntactic analysis of the different components of an MWE.

Specific MWE categories. Specific MWE categories such as multiword named entities or multiword terms pose particular challenges to MT systems, because they may require specific translations not directly deducible from the translations of their components. These categories are very productive, that is, new multiword named entities and terms are constantly being created, so it is difficult to have complete and updated lexical resources for use during the translation process. For term identification, bilingual or multilingual term glossaries might be applied to transform terms into an intermediate representation (words concatenated by underscore). But when these resources are missing, for new domains or for under-resourced languages, multiword named entities and multiword terms can be annotated as a single token with the help of specific techniques for named entity recognition (Tan and Pal 2014), or term extraction (Bouamor, Semmar, and Zweigenbaum 2012b) designed for monolingual, parallel, or comparable data (Morin and Daille 2010).

Closed compounds, obtained by concatenating several lexemes with any parts of speech, are typical of Germanic languages and represent another difficult task for MT. This category of expressions can be lexicalized, that is, they belong to the lexicon of a language as a single meaning unit, such as the German word *Schwiegereltern* (*parents-in-law*) or non-lexicalized, that is, the individual words keep their meanings when combined, for instance, the German neologism *Helikoptereltern* (*helicopter parents*). They are usually translated into several target language words. Their meaning might be more or less compositional. MT systems fail to correctly translate these compounds because of their low frequencies and their variability. Moreover, non-compositional compounds have unpredictable meaning.

Splitting strategies can be applied to cut the compounds into subsequent words to improve translation quality (Fritzinger and Fraser 2010; Stymne, Cancedda, and Ahrenberg 2013). Splitting is done by identifying component words in the corpus or by prefix and suffix identification together with distributional semantics (Weller et al. 2014) or by using a morphosyntactic tagger and parser (Cap et al. 2014). Oversplitting can also be a problem: Splitting non-compositional compounds may generate erroneous translations. Some methods aim to distinguish between compositional and non-compositional compounds and split only the compositional ones (Weller et al. 2014). A postprocessing step is required to merge components back into compounds once a translation is generated using a system trained on split compounds. Some methods replace the compounds by paraphrases (Ullman and Nivre 2014) before translating them.

Preprocessing methods (concatenation or decomposition) tag MWEs in the input data: This strategy is effective for SMT or for RBMT and avoids data sparsity. As a drawback, such methods can only handle contiguous MWEs. Also, without filtering, MWE identification methods can add noise into the MT system by annotating candidates which are not really MWEs. This results in MT performance loss.

5.2.3 Identification During MT. MWE-aware strategies can be best differentiated according to the MT paradigm under discussion.

MWE-aware strategies in SMT. Phrase-based SMT systems build probabilistic models during the training phase, based on simple word alignment and on a phrase table (a list of pairs of *n*-grams, their *n*-gram translation, and their translation scores). Then, the core translation process is ultimately determined by the contents of the phrase table—thus, one way to regard MWE identification during MT is in terms of changing the contents of the phrase table adaptively, during the training phrase (Carpuat and Diab 2010). Several observed approaches are: (1) changing the training data dynamically (word alignment or the parallel corpus) to take into account MWEs and then retraining the system; (2) modifying the phrase table directly by including information about MWEs and their translations. In both strategies, the use of MWE identification and discovery tools is essential to improve the quality of the translation.

Modifying training data. A frequent strategy completes simple word alignment with many-to-many, many-to-one, or one-to-many alignments to solve translation asymmetries (Melamed 1997; Carpuat and Diab 2010; Okita 2012). Word alignment completion is based on simple word alignment and on MWE identification tools, designed for specific MWE categories (Tan and Pal [2014] for multiword named entities; Bouamor, Semmar, and Zweigenbaum [2012b] and Okita and Way [2011] for terms; Ramisch, Villavicencio, and Boitet [2010] for general MWEs). Alternatively, MWE identification and alignment is performed using bilingual lexical resources, with translation alongside an *n*-gram language model to help with disambiguation (Bungum et al. 2013). The resulting many-to-many word alignment is used to retrain the system in order to build a new phrase table. As a consequence, the phrase table takes into account MWEs and their translations.

Alternatively, bilingual dictionaries of MWEs are added as additional training data to the parallel corpus (Babych and Hartley 2010; Tan and Pal 2014).

Modifying the phrase table. Usually, a bilingual list of MWEs and their equivalents is dynamically extracted from the simple word alignment using specific MWE discovery tools (Bouamor, Semmar, and Zweigenbaum 2012b; Kordoni and Simova 2012; Pal, Naskar, and Bandyopadhyay 2013). Then, the phrase table is completed with the bilingual lists of MWEs and the probabilities are modified accordingly (Lambert and Banchs 2005) or added into a new phrase table with the probability set to 1 (Ren et al. 2009).

An alternate strategy consists of adding new features in the phrase table, such as the number of MWEs present in the bilingual aligned phrases (Carpuat and Diab 2010) or the property that the parallel phrase contains a bilingual MWE (Ren et al. 2009). In this way, the translation quality is improved for certain specific MWE categories or languages (Costa-Jussà, Daudaravicius, and Banchs 2010). The modified phrase table contains, indeed, the correct translations of MWEs, thus avoiding an incorrect word-for-word translation during the decoding phase and helping disambiguation.

More complex models are proposed in syntax-based SMT (Na et al. 2010) or in hierarchical SMT (Chiang 2007). These approaches use grammars to handle discontiguous components and find their translation directly: parsing improves the translation process

(according to BLEU and METEOR scores) by providing trees and transfer rules based on parsed data (Wei and Xu 2011).

MWE-aware strategies in EBMT and RBMT. EBMT (Gangadharaiah, Brown, and Carbonell 2006) or RBMT strategies (Anastasiou 2008; Forcada et al. 2011; Monti et al. 2011) dynamically apply rules to handle MWE translations. Some rules are identified from the syntactic tree alignments (Segura and Prince 2011) and integrated into an EBMT system to handle discontiguous MWEs.

RBMT systems use large lexicons to handle contiguous MWEs and apply the correct translation strategy: a simple word-for-word translation strategy or a compositional rule (Wehrli et al. 2009). Discontiguous MWEs are identified using parsing output or some linguistic patterns. Several RBMT systems identify MWEs and generate translations on the basis of formal representations of natural language texts such as parse trees (Wehrli et al. 2009) or intermediate representation languages like minimal recursion semantics (Oepen et al. 2004), a semantico-syntactic abstraction language (Monti et al. 2011; Barreiro et al. 2013). Transfer rules handle MWE variability and discontiguity (Forcada et al. 2011) and are manually defined or automatically learned from parallel corpora (Haugereid and Bond 2011).

Discontiguous or variable MWEs represent an important source of translation errors. These methods have the advantage of handling discontiguous or variable MWEs with the help of rules for RBMT or by completing word alignments dynamically in SMT.

5.3 Evaluation of MWE-Aware MT

The evaluation of MWEs translation quality remains an open challenge, whatever MT paradigm is adopted (Monti et al. 2012; Ramisch, Besacier, and Kobzar 2013; Barreiro et al. 2014), because of a lack of shared assessment methodologies, benchmarking resources, and annotation guidelines.

With reference to the assessment methodologies, automatic evaluation metrics such as BLEU (Papineni et al. 2002) do not specifically take MWE translation quality into account. For instance, BLEU is based on shared words between the candidate and the reference translation, and gives only a very general indication about quality. Thus, it cannot be considered as a suitable metric for the kind of more differentiated analysis required to identify specific gaps in the coverage of the system, as is needed for MWEs. There have been a few attempts to adapt automatic evaluation metrics towards a more fine-grained MT error analysis (Babych and Hartley 2010; Stymne, Cancedda, and Ahrenberg 2013; Salehi et al. 2015). Extrinsic evaluations in MT have also been performed, mainly for SMT. For instance, Carpuat and Diab (2010) conducted a pilot study for a task-oriented evaluation of MWE translation in SMT, whereas Bouamor, Semmar, and Zweigenbaum (2012a) consider SMT as an extrinsic evaluation of the usefulness of automatically discovered MWEs and explore strategies for integrating them in a SMT system, aiming at a more thorough error analysis of MWE translation.

Another important drawback in this field is represented by the fact that parallel corpora annotated with MWEs, which are important and necessary gold standard resources for the evaluation of MT translation quality, are very scarce. MWE annotation is indeed a complex and time-consuming task. Annotated resources are usually produced manually and require a large number of experts. In addition, annotating MWEs in parallel corpora requires the correct delimitation of MWEs (contiguous vs. discontiguous expressions) to classify and to disambiguate them and to handle not only the multilingual dimension, but also translation asymmetries between languages (Section 5.1). Moreover, each category of MWEs has its own set of properties.

MWE-annotated benchmarking resources useful for translation quality evaluation are usually available for (1) specific MWE categories, (2) specific language pairs, (3) a specific MWE alignment tool or integration strategy in MT systems, or (4) specific approaches to handling MWEs in MT. Evaluation data consist mainly of small parallel corpora, manually built by carefully selecting sentences containing specific categories of MWE to avoid data sparseness, aligned either with human translations collected from the Web or generated by commercial MT systems (Google Translate, Bing, OpenLogos). Previous work that makes use of such resources includes Ramisch, Besacier, and Kobzar (2013) for verb-particle constructions in English and French; Barreiro et al. (2013) for different categories of MWE in language pairs involving English to French, Italian, and Portuguese; Laporte (2014) for French-Romanian verb-noun idioms and collocations; Weller et al. (2014) for compositional noun compounds, compositional verb compounds, and a set of non-compositional compounds in German–English; Barreiro et al. (2014) for light-verb constructions in English to Italian, French, Portuguese, German, and Spanish; and Schottmüller and Nivre (2014) for verb-particle constructions in the German–English language pair. In addition, these linguistic resources are annotated only with a limited set of MWE categories such as, for instance, light-verb constructions (Vincze 2012; Rácz, Nagy, and Vincze 2014). They are of variable granularity, so some annotation schemes consider only MWE categories, whereas others include additional information such as POS and degree of fixedness.

There are only very few instances of parallel corpora annotated with several categories of MWEs and with different types of correspondences (many-to-one, one-to-many, and many-to-many translations), such as those created by Monti, Sangati, and Arcan (2015), Tutin et al. (2015), and Flickinger et al. (2012). Moreover, the lack of homogeneity represents a real obstacle to the effective reuse of existing annotated data.

Concerning MWE annotation guidelines, only very few papers describe the procedures adopted during resource development. Comprehensive approaches to MWE annotation in parallel corpora, that is, which take into account a large inventory of MWE categories, include Monti, Sangati, and Arcan (2015), who developed a parallel English–Italian corpus, and Tutin et al. (2015), who worked on the French part of a parallel French–English corpus.

In conclusion, the evaluation of MWE processing in MT is still an open issue, as we will discuss in the next section.

5.4 Open Issues in MWE-Aware MT

Orchestration is an open issue for MT. Existing systems have adopted rather specific strategies for handling MWEs. Thus only a few categories of MWE are addressed, namely verbal expressions (such as light-verb constructions, verb-noun collocations, and verb-particle constructions), multiword terms, multiword named entities, or noun compounds for specific domains or languages.

For some categories (e.g., terms, multiword named entities), preprocessing methods seem more effective, whereas for others, such as collocations, MWE identification during the MT process improves the results. The generalization of procedures to uniformly handle other categories of MWE (multiword adverbials, nominal compounds) has not been investigated.

In short, some orchestration strategies (identification before or during translation) seem to be more effective for certain MWE categories, but this depends on their properties (variability, non-compositionality) and on the availability of resources for the domain and languages involved.

Manually built resources are not available for all language pairs and domains, so MWE discovery techniques extend MT data with bilingual MWE lists, but need large quantities of parallel data. Indeed, building bilingual MWE lists requires a combination of parallel data and of MWE discovery tools based on statistical measures and on linguistic information. Lexical resources are crucial for the identification of non-compositional expressions: the translation is done using an intermediate semantic (conceptual) representation, but these resources again have limited coverage for specific domains or languages. The main difficulty is to propose generic representations for such lexical resources and generic methods to create new resources for several languages. An open challenge is how to create lexical resources for under-resourced languages by exploiting comparable data, monolingual resources, or domain specificity.

Some phenomena are not yet fully exploited by MWE identification or discovery methods. For example, ambiguity in translation (both literal translation and non-compositional translation are available) might be used by these methods to find new MWE candidates.

In the previous section, we have shown that MWE-aware MT evaluation requires further research. Detailed guidelines are lacking, as are parallel corpora containing a significant number of annotated MWE examples. Some combination of human expertise with the output of MT systems might help solve the problem of creating larger evaluation resources. In addition, only few contributions have been devoted to the comparison and evaluation of different MT paradigms regarding MWEs. A systematic investigation in this respect for the different MWEs and languages may help in identifying the most suitable approach to the translation of specific categories of MWEs.

6. Conclusions and Open Issues

In this survey, we have presented a conceptual framework for MWE processing that facilitates a clear understanding of what MWE processing is, and that delineates its subtasks and their subsequent interactions with use cases such as parsing and MT. It allows us to draw several conclusions about MWE processing and its interactions with the selected use cases.

MWE properties: challenges and opportunities. The basic characteristics of MWEs are such that they present both challenges and opportunities. Discontiguity, for example, is clearly a challenge for both discovery and identification, but an opportunity for parsing to address this particular issue by providing a syntactic analysis linking non-adjacent words.

Interactions between tasks and use cases. The separation between the subtasks of MWE processing and use cases is a key feature of our framework and allows us to explore the interdependencies between the two. These take the form of support (i.e., *helping*) relations that can work in either direction, depending on whether the MWE properties at stake are a challenge or an opportunity for the task under consideration. Parsing can support identification, for example, to join discontiguous elements in an MWE by means of syntactic relations; conversely, identification can help parsing in legitimating a special treatment of identified MWEs by the parser.

Orchestration. With respect to the way the several tasks (discovery, identification, parsing, and MT) are orchestrated, we can conclude that the direction of the support relations found can determine which orchestration scenarios are most suitable. For example, in order for identification methods to profit from syntactic analysis, they must be preceded by syntactic analysis.

MWE categories. Cutting across these dimensions is the category of MWE one is trying to process. Because characteristics of MWEs are category-dependent (e.g., verbal MWEs are often discontiguous whereas compounds are not), category affects the direction of support relations and hence the choice of orchestration strategy.

As such, the framework defines three dimensions for positioning previous work in this area: the tasks (identification, discovery, parsing, and MT), the orchestration scenarios (before, after, during), and the defining characteristics of different MWE categories. Within the space defined by these dimensions, some regions are densely populated, whereas others are sparse.

In some cases, the reasons for sparseness (i.e., little research carried out) are obvious. For example, MWE categories that do not exhibit the characteristic of discontiguity benefit less from parsing and therefore multiword named entities and compounds are often dealt with by preprocessing and not during or after parsing/MT.

In other cases, less-populated areas might represent promising avenues for future work. For example, we have mentioned how parsing can support discovery for the property of discontiguity. However, there is little evidence for the inverse relation: Few studies have looked at how the results of discovery, in the form of a lexicon or perhaps rules, might support parsing.

Apart from pointers as to where more research would be needed in the area provided by the framework, we found the following issues for MWE processing to be most pressing:

Evaluation. Both the two subtasks of MWE processing (discovery and identification) individually, as well as MWE-aware parsing and MT, have several open issues with respect to evaluation. In the case of discovery, evaluation is usually done intrinsically by means of expert judgments, gold standards, and dedicated test sets, but these lack the scale, naturalness, and coverage of possible extrinsic evaluation in downstream applications. Evaluation of identification, parsing, and MT are negatively affected by inconsistencies in annotated data. In addition, annotations should include finer-grained properties such as embeddings (for parsing). For MT, awareness of the language/domain-specificity in MWE representations could help advance MWE evaluation. Evaluation metrics need to be further developed as well as guidelines for the annotation of MWEs.

Large-scale, comparative evaluations. Apart from the fact that evaluation is an open issue for each of the processing tasks and the two use cases individually, large-scale comparative evaluations, which contrast parameters across the three dimensions discussed (task at hand, orchestration scenario, and MWE category), are badly needed to gain further insight into the optimum handling of particular regions of the space. Although the sections on parsing and MT position previous work within the framework and provide detailed overviews of the findings, distribution over the multidimensional space is uneven and few comparative evaluations have been carried out.

Additional use cases. In this article, we have started to analyze the relationship between MWE processing and NLP applications by looking at two particular use cases (parsing and MT), but consideration of other use cases might favor particular strategies. For example, information extraction from news articles might rely on relatively fixed multiword named entities, implying that pre-identification is optimal. It might well be different for sentiment analysis within political discourse. Clearly, more research is necessary to establish such relationships.

Coverage. As input to identification, hand-built dictionaries are still the main resource, though they are known to be limited in terms of coverage. Coverage is not only limited to certain domains and languages, but to currency, because new terms

are created on a regular basis. Automatic discovery methods can be designed to pick these up from corpora on a continuous basis and language-independent automatic discovery methods can build resources for several languages. The use of such automatic discovery methods in downstream applications, such as identification and subsequently in parsing and MT, is therefore a promising path to better coverage. In doing so, MWE discovery would gain the extrinsic evaluation methods needed.

Appendix A. MWE Resources and Tools

This appendix lists some outstanding MWE resources and tools, among many more.

Type	Name	Reference
List of resources	MWE SIGLEX Section Web site	Grégoire, Evert, and Krenn (2008)
List of resources	PARSEME list of resources	Losnegaard et al. (2016)
MWE-aware corpus	Streusle	Schneider et al. (2014b)
MWE-aware corpus	Wiki50	Vincze, Nagy, and Berend (2011)
MWE-aware corpus	PARSEME shared task corpora	Savary et al. (2017)
MWE-aware corpus	TED-MWE	Monti, Sangati, and Arcan (2015)
MWE-aware lexicon	BabelNet - multilingual	Navigli and Ponzetto (2012)
MWE-aware lexicon	JRC-Names - multilingual	Ehrmann, Jacquet, and Steinberger (2017)
MWE-aware treebank	IMST Turkish Treebank	Adalı et al. (2016)
MWE-aware treebank	IWT Turkish Treebank	Adalı et al. (2016)
MWE-aware treebank	Eukalyptus Treebank of Written Swedish	Adesam, Bouma, and Johansson (2015)
MWE-aware treebank	Latvian Treebank	Pretkalnia and Rituma (2012)
MWE-aware treebank	Prague Dependency Treebank	Bejček et al. (2012)
MWE-aware treebank	RoRefTrees Romanian Treebank	Mititelu and Irimia (2015)
MWE-aware treebank	French Treebank	Abeillé, Clément, and Toussenel (2003)
MWE-aware treebank	Szeged Hungarian Treebank	Vincze et al. (2010)
Discovery and identification tool	mwetoolkit	Ramisch (2015)
Discovery tool	Varro toolkit	Martens and Vandeghinste (2010)
Discovery tool	Text:NSP	Banerjee and Pedersen (2003)
Discovery tool	UCS	Evert (2005)
Discovery tool	LocalMaxs	da Silva et al. (1999)
Discovery tool	Druid	Riedl and Biemann (2015)
Identification tool	jMWE	Finlayson and Kulkarni (2011)
Identification tool	AMALGrAM	Schneider et al. (2014a)
Identification tool	LGTagger	Constant and Sigogne (2011)
MWE-aware parser	Stanford parser in French	Green, de Marneffe, and Manning (2013)
MWE-aware MT	ITS-2	Wehrli et al. (2009)
MWE-aware MT	Apertium	Forcada et al. (2011)
MWE-aware MT	OpenLogos	Barreiro et al. (2011)

Acknowledgments

This work has been supported by the PARSEME project (Cost Action IC1207). Special thanks to Federico Sangati, who not only provided valuable suggestions, references, and feedback, but also contributed important initial written material to the foundations of this survey. We would like to thank all members of the PARSEME working group 3 who contributed suggestions and references: Jan Genci, Tunga Güngör, Tomas Krilavicius, Justina Mandravickaite, Michael Oakes, Yannick Parmentier, Carla Parra Escartín, Gerold Schneider, Inguna Skadiņa, Dan Tufis, Éric Villemonte de la Clergerie, Veronika Vincze, and Eric Wehrli. This work has been partially funded by the French National Research Agency (ANR) through the PARSEME-FR project (ANR-14-CERA-0001) and by a TUBITAK 1001 grant (no: 112E276).

References

Abeillé, Anne. 1995. The flexibility of French idioms: A representation with lexicalised tree adjoining grammar. In Everaert, M., van der Linden, E-J., Schenk, A., Schreuder, R., editors, *Idioms: Structural and Psychological Perspectives*. Lawrence Erlbaum Associates, New Jersey, chapter 1, pages 15–41.

Abeillé, Anne, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. In *Treebanks*. Springer, Berlin, 165–187.

Abeillé, Anne and Yves Schabes. 1989. Parsing idioms in lexicalized TAGs. In *Proceedings of EACL 1989*, pages 1–9, Manchester.

Adalı, Kübra, Tutkum Dinç, Memduh Gokirmak, and Gülşen Eryiğit. 2016. Comprehensive annotation of multiword expressions for Turkish. In *Proceedings of TurCLing 2016 at CICLING 2016*, pages 60–66, Konya.

Adesam, Yvonne, Gerlof Bouma, and Richard Johansson. 2015. Multiwords, word senses and multiword senses in the Eukalyptus treebank of written Swedish. In *Proceedings of TLT 2014*, page 3.

Ait-Mokhtar, Salah, Jean-Pierre Chanod, and Claude Roux. 2002. Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*, 8(2-3):121–144.

Anastasiou, Dimitra. 2008. Identification of idioms by machine translation: a hybrid research system vs. three commercial systems. In *Proceedings of EAMT 2008*, pages 12–20, Hamburg.

Anastasiou, Dimitra, Chikara Hashimoto, Preslav Nakov, and Su Nam Kim, editors. 2009. *Proceedings of the ACL 2009 Workshop on MWEs*. ACL, Singapore.

Arun, Abhishek and Frank Keller. 2005. Lexicalization in crosslinguistic probabilistic parsing: The case of French. In *Proceedings of ACL 2005*, pages 306–313, Ann Arbor, MI.

Attia, Mohammed A. 2006. Accommodating multiword expressions in an Arabic LFG grammar. In *Proceedings of FinTAL 2006*, pages 87–98, Turku.

Babych, Bogdan and Anthony Hartley. 2010. Automated error analysis for multiword expressions: Using BLEU-type scores for automatic discovery of potential translation errors. *Evaluation of Translation Technology*, 8:81–104.

Baldwin, Timothy. 2005. Deep lexical acquisition of verb-particle constructions. *Computer Speech & Language*, 19(4):398–414.

Baldwin, Timothy, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on MWEs*, pages 89–96, Sapporo.

Baldwin, Timothy and Su Nam Kim. 2010, Multiword expressions. In *Handbook of Natural Language Processing*, 2nd edition. CRC Press, Taylor and Francis Group, Boca Raton, FL, pages 267–292.

Banerjee, Satanjeev and Ted Pedersen. 2003. The design, implementation, and use of the ngram statistics package. In *Proceedings of CICLing 2003*, pages 370–381, Mexico City.

Bannard, Colin. 2002. Statistical techniques for automatically inferring the semantics of verb-particle constructions. Lingo Working Paper No. 2002-06, University of Edinburgh.

Bannard, Colin. 2005. Learning about the meaning of verb-particle constructions from corpora. *CSL Special Issue on MWEs*, 19(4):467–478.

Bansal, Mohit and Dan Klein. 2011. Web-scale features for full-scale parsing. In *Proceedings of ACL 2011*, pages 693–702, Portland, OR.

Baptista, Jorge, Graça Fernandes, Rui Talhadas, Francisco Dias, and Nuno Mamede. 2015. Implementing European Portuguese verbal idioms in a natural language processing system. In *Proceedings of Europhras 2015*. Málaga, pages 102–115.

Barreiro, Anabela. 2008. *Make it Simple with Paraphrases: Automated Paraphrasing for*

Authoring Aids and Machine Translation. Ph.D. thesis, Universidate do Porto, Porto.
Barreiro, Anabela, Johanna Monti, Brigitte Orliac, and Fernando Batista. 2013. When multiwords go bad in machine translation. In *MT Summit Workshop Proceedings on Multi-word Units in Machine Translation and Translation Technology*, pages 26–33, Nice.
Barreiro, Anabela, Johanna Monti, Brigitte Orliac, Susanne Preuß, Kutz Arrieta, Wang Ling, Fernando Batista, and Isabel Trancoso. 2014. Linguistic evaluation of support verb constructions by OpenLogos and Google Translate. In *Proceedings of LREC 2014*, pages 35–40, Reykjavik.
Barreiro, Anabela, B. Scott, W. Kasper, and Bernd Kiefer. 2011. Openlogos machine translation: Philosophy, model, resources and customization. *Machine Translation*, 25:107–126.
Bejček, Eduard, Jarmila Panevová, Jan Popelka, Pavel Straňák, Magda Ševčíková, Jan Štěpánek, and Zdeněk Žabokrtský. 2012. Prague dependency treebank 2.5 a revisited version of PDT 2.0. In *Proceedings of COLING 2012*, pages 231–246, Mumbai.
Birke, Julia and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of EACL 2006*, pages 329–336, Trento.
Black, Ezra, Steve Abney, Dan Flickinger, Claudia Gdaniec, Ralph Grishman, Phil Harrison, Don Hindle, Robert Ingria, Fred Jelinek, Judith Klavans, Mark Liberman, Mitchell Marcus, Salim Roukos, Beatrice Santorini, and Tomek Strzalkowski. 1991. Procedure for quantitatively comparing the syntactic coverage of English grammars. In *Proceedings of the Workshop on Speech and Natural Language*, pages 306–311, Pacific Grove, CA.
Blunsom, Phil and Timothy Baldwin. 2006. Multilingual deep lexical acquisition for HPSGs via supertagging. In *Proceedings of EMNLP 2006*, pages 164–171, Sydney.
Bond, Francis, Su Nam Kim, Preslav Nakov, and Stan Szpakowicz, editors. 2013. *Natural Language Engineering Special Issue on Noun Compounds*, volume 19(3). Cambridge Univ. Press, Cambridge.
Bond, Francis, Anna Korhonen, Diana McCarthy, and Aline Villavicencio, editors. 2003. *Proceedings of ACL 2003 Workshop on MWEs*, ACL, Sapporo.
Bonial, Claire, Julia Bonn, Kathryn Conger, Jena D. Hwang, and Martha Palmer. 2014. Propbank: Semantics of new predicate types. In *Proceedings of LREC 2014*, pages 3013–3019, Reykjavik.
Bouamor, Dhouha, Nasredine Semmar, and Pierre Zweigenbaum. 2012a. Automatic construction of a multiword expressions bilingual lexicon: A statistical machine translation evaluation perspective. In *Proceedings of COLING 2012*, pages 95–107, Bombay.
Bouamor, Dhouha, Nasredine Semmar, and Pierre Zweigenbaum. 2012b. Identifying bilingual multi-word expressions for statistical machine translation. In *Proceedings of LREC 2012*, pages 674–679, Istanbul.
Boukobza, Ram and Ari Rappoport. 2009. Multi-word expression identification using sentence surface features. In *Proceedings of EMNLP 2009*, pages 468–477, Singapore.
Boullier, Pierre and Benoît Sagot. 2005. Efficient and robust LFG parsing: SxLfg. In *Proceedings of IWPT 2005*, pages 1–10, Vancouver.
Breidt, Elisabeth, Frédérique Segond, and Giuseppe Valetto. 1996. Formal description of multi-word lexemes with the finite-state formalism IDAREX. In *Proceedings of COLING 1996*, pages 1036–1040, Copenhagen.
Brun, Caroline. 1998. Terminology Finite-state preprocessing for computational LFG. In *Proceedings of COLING*, pages 196–200, Montreal.
Buchholz, Sabine and Erwin Marsi. 2006. Conll-X shared task on multilingual dependency parsing. In *Proceedings of CoNLL 2006*, pages 149–164, New York, NY.
Bungum, Lars, Björn Gambäck, André Lynum, and Erwin Marsi. 2013. Improving word translation disambiguation by capturing multiword expressions with dictionaries. In *Proceedings of the NAACL/HLT 2013 Workshop on MWEs*, pages 21–30, Atlanta, GA.
Cafferkey, Conor, Deirdre Hogan, and Josef van Genabith. 2007. Multi-word units in treebank-based probabilistic parsing and generation. In *Proceedings of RANLP 2007*, pages 98–103, Borovets.
Cahill, Aoife. 2004. *Parsing with Automatically Acquired, Wide-coverage, Robust, Probabilistic LFG Approximations*. Ph.D. thesis, Dublin City University School of Computing.
Calzolari, Nicoletta, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine Macleod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of LREC 2002*, pages 1934–1940, Canary Islands.

Candito, Marie and Matthieu Constant. 2014. Strategies for contiguous multiword expression analysis and dependency parsing. In *Proceedings of ACL 2014*, pages 743–753, Baltimore, MD.

Cap, Fabienne, Alexander Fraser, Marion Weller, and Aoife Cahill. 2014. How to produce unseen teddy bears: Improved morphological processing of compounds in SMT. In *Proceedings of EACL 2014*, pages 579–587, Gothenburg.

Cap, Fabienne, Manju Nirmal, Marion Weller, and Sabine Schulte im Walde. 2015. How to account for idiomatic German support verb constructions in statistical machine translation. In *Proceedings of NAACL-HLT 2015*, pages 19–28, Denver, CO.

Carpuat, Marine and Mona Diab. 2010. Task-based evaluation of multiword expressions: A pilot study in statistical machine translation. In *Proceedings of NAACL/HLT 2010*, pages 242–245, Los Angeles, CA.

Caseli, Helena de Medeiros, Carlos Ramisch, Maria das Gracas Volpe Nune, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 44(1-2):59–77.

Chanod, Jean Pierre and Pasi Tapanainen. 1996. A non-deterministic tokeniser for finite-state parsing. In *Proceedings of the ECAI 1996 Workshop on Extended Finite State Models of Language*, pages 10–12, Budapest.

Charniak, Eugene and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of ACL 2005*, pages 173–180, Ann Arbor, MI.

Chen, Danqi and Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of EMNLP 2014*, pages 740–750, Doha.

Chiang, David. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33 (2):202–228.

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of EMNLP 2014*, pages 1724–1734, Doha.

Choueka, Yaacov. 1988. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. *Proceedings of RIA*, pages 609–624, Cambridge, MA.

Church, Kenneth Ward and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Constant, Matthieu, Marie Candito, and Djamé Seddah. 2013. The LIGM-Alpage architecture for the SPMRL 2013 shared task: Multiword expression analysis and dependency parsing. In *Proceedings of SPRML 2013*, pages 46–52, Seattle, WA.

Constant, Matthieu, Joseph Le Roux, and Anthony Sigogne. 2013. Combining compound recognition and PCFG-LA parsing with word lattices and conditional random fields. *ACM TSLP Special Issue on MWEs*, 10(3):1–24.

Constant, Matthieu, Joseph Le Roux, and Nadi Tomeh. 2016. Deep lexical segmentation and syntactic parsing in the easy-first dependency framework. In *Proceedings of NAACL/HLT 2016*, pages 1095–1101, San Diego, CA.

Constant, Matthieu and Joakim Nivre. 2016. A transition-based system for joint lexical and syntactic analysis. In *Proceedings of ACL 2016*, pages 161–171, Berlin.

Constant, Matthieu and Anthony Sigogne. 2011. MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proceedings of the ACL 2011 Workshop on MWEs*, pages 49–56, Portland, OR.

Constant, Matthieu, Anthony Sigogne, and Patrick Watrin. 2012. Discriminative strategies to integrate multiword expression recognition and parsing. In *Proceedings of ACL 2012*, pages 204–212, Jeju Island.

Cook, Paul, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the ACL 2007 Workshop on MWEs*, page 41–48, Prague.

Copestake, Ann, Fabre Lambeau, Aline Villavicencio, Francis Bond, Timothy Baldwin, Ivan A. Sag, and Dan Flickinger. 2002. Multiword expressions: Linguistic precision and reusability. In *Proceedings of LREC 2002*, page 1941–1947, Canary Islands.

Cordeiro, Silvio, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2016. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of ACL 2016*, pages 1986–1997, Berlin.

Costa-Jussà, Marta, R., Vidas Daudaravicius, and Rafael E. Banchs. 2010. Integration of

statistical collocation segmentations in a phrase-based statistical machine translation system. In *Proceedings of EAMT 2010*, Saint-Raphaël.
Diab, Mona and Pravin Bhutada. 2009. Verb noun construction MWE token classification. In *Proceedings of the ACL 2009 Workshop on MWEs*, pages 17–22, Singapore.
Doucet, Antoine and Helana Ahonen-Myka. 2004. Non-contiguous word sequences for information retrieval. In *Proceedings of the ACL 2004 Workshop on MWEs*, pages 88–95, Barcelona.
Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61–74.
Durrett, Greg and Dan Klein. 2015. Neural CRF parsing. In *Proceedings of ACL 2015*, pages 302–312, Beijing.
Dyer, Chris, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *Proceedings of ACL 2015*, pages 334–343, Beijing.
Ehrmann, Maud, Guillaume Jacquet, and Ralf Steinberger. 2017. JRC-names: Multilingual entity name variants and titles as linked data. *Semantic Web*, 8(2):283–295.
Eryiğit, Gülşen, Tugay İlbay, and Ozan Arkan Can. 2011. Multiword expressions in statistical dependency parsing. In *Proceedings of SPMRL 2011*, pages 45–55, Dublin.
Evert, Stefan. 2005. *The Statistics of Word Co-occurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart, Stuttgart.
Evert, Stefan. 2009. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, volume 2. Mouton de Gruyter, pages 1212–1248.
Farahmand, Meghdad and James Henderson. 2016. Modeling the non-substitutability of multiword expressions with distributional semantics and a log-linear model. In *Proceedings of the ACL 2016 Workshop on MWEs*, pages 61–66, Berlin.
Fazly, Afsaneh, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.
Fazly, Afsaneh and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of EACL 2006*, pages 337–344, Trento.
Fillmore, Charles J., Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64:501–538.
Finkel, Jenny Rose and Christopher D. Manning. 2009. Joint parsing and named entity recognition. In *Proceedings of NAACL 2009*, pages 326–334, Boulder, CO.
Finlayson, Mark and Nidhi Kulkarni. 2011. Detecting multi-word expressions improves word sense disambiguation. In *Proceedings of the ACL 2011 Workshop on MWEs*, pages 20–24, Portland, OR.
Flickinger, Daniel, Valia Kordoni, Yi Zhang, António Branco, Kiril Simov, Petya Osenova, Catarina Carvalheiro, Petya Osenova, Catarina Carvalheiro, Francisco Costa, and Sérgio Castro. 2012. Pardeepbank: Multiple parallel deep treebanking. In *Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories*, pages 97–108.
Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: A free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.
Fothergill, Richard and Timothy Baldwin. 2012. Combining resources for MWE-token classification. In *Proceedings of *SEM 2012*, pages 100–104, Montréal.
Fritzinger, Fabienne and Alexander Fraser. 2010. How to avoid burning ducks: Combining linguistic analysis and corpus statistics for German compound processing. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, WMT 2010*, pages 224–234, Uppsala.
Gangadharaiah, Rashmi, Ralf Brown, and Jaime Carbonell. 2006. Spectral clustering for example based machine translation. In *Proceedings of NAACL/HLT 2006*, pages 41–44, New York, NY.
Ghoneim, Mahmoud and Mona Diab. 2013. Multiword expressions in the context of statistical machine translation. In *Proceedings of IJCNLP 2013*, pages 1181–1187, Nagoya.
Girju, Roxana, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the semantics

of noun compounds. *CSL Special Issue on MWEs*, 19(4):479–496.
Green, Spence, Marie-Catherine de Marneffe, John Bauer, and Christopher D. Manning. 2011. Multiword expression identification with tree substitution grammars: A parsing tour de force with French. In *Proceedings of EMNLP 2011*, pages 725–735, Edinburgh.
Green, Spence, Marie-Catherine de Marneffe, and Christopher D. Manning. 2013. Parsing models for identifying multiword expressions. *Computational Linguistics*, 39(1):195–227.
Grégoire, Nicole, Stefan Evert, and Brigitte Krenn, editors. 2008. *Proceedings of the LREC 2008 Workshop Towards a Shared Task for MWEs (MWE 2008)*. Marrakech.
Gross, Maurice. 1989, The use of finite automata in the lexical representation of natural language. In Maurice Gross and Dominique Perrin, editors, *Electronic Dictionaries and Automata in Computational Linguistics: LITP Spring School on Theoretical Computer Science Saint-Pierre d'Oléron, France, May 25–29, 1987 Proceedings*. Springer Berlin Heidelberg, Berlin, Heidelberg, page 34–50.
Hashimoto, Chikara and Daisuke Kawahara. 2008. Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features. In *Proceedings of EMNLP 2008*, pages 992–1001, Waikiki, HI.
Hashimoto, Chikara, Satoshi Sato, and Takehito Utsuro. 2006. Japanese idiom recognition: Drawing a line between literal and idiomatic meanings. In *Proceedings of COLING/ACL 2006*, pages 353–360, Sydney.
Haugereid, Petter and Francis Bond. 2011. Extracting transfer rules for multiword expressions from parallel corpora. In *Proceedings of the ACL 2011 Workshop on MWEs*, pages 92–100, Portland, OR.
Hoang, Hieu and Philipp Koehn. 2010. Improved translation with source syntax labels. In *Proceedings of the ACL 2010 Workshop on Statistical Machine Translation and Metrics MATR*, pages 409–417, Uppsala.
Hoang, Hung Huu, Su Nam Kim, and Min-Yen Kan. 2009. A re-examination of lexical association measures. In *Proceedings of the ACL 2009 Workshop on MWEs*, pages 31–39, Singapore.
Hollenstein, Nora, Nathan Schneider, and Bonnie Webber. 2016. Inconsistency detection in semantic annotation. In *Proceedings of LREC 2016*, pages 3986–3990, Portorož.
Jagfeld, Glorianna and Lonneke van der Plas. 2015. Towards a better semantic role labelling of complex predicates. In *Proceedings of NAACL Student Research Workshop*, pages 33–39, Denver, CO.
Joshi, Aravind K. and Phil Hopeli. 1996. A parser from antiquity. *Natural Language Engineering*, 2:291–294.
Joshi, Aravind K., Leon S. Levy, and Masako Takahashi. 1975. Tree Adjunct Grammars. *Journal of Computer and System Sciences*, 10(1):136–163.
Justeson, John S. and Slava M. Katz. 1995. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27.
Kalchbrenner, Nal and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of EMNLP 2013*, pages 1700–1709.
Kaplan, Ronald M. 1989. The formal architecture of lexical-functional grammar. *Journal of Informations Science and Engineering*, 5(4):305–322.
Katz, Graham and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL/COLING 2006 Workshop on MWEs*, pages 12–19, Sydney.
Keller, Frank and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.
Kim, Jae Dong, Ralf Brown, and Jaime G. Carbonell. 2010. Chunk-based EBMT. In *Proceedings of EAMT 2010*, Saint Raphael.
Koehn, Philipp. 2010. *Statistical Machine Translation*, 1st edition. Cambridge University Press, New York, NY.
Koehn, Philipp and Hieu Hoang. 2007. Factored translation models. In *Proceedings of EMNLP/CoNLL 2007*, pages 868–876, Prague.
Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL 2003*, pages 48–54, Edmonton.
Kong, Lingpeng, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. Dependency parsing for tweets. In *Proceedings of EMNLP 2014*, pages 1001–1012, Doha.
Kordoni, Valia and Iva Simova. 2012. Multiword expressions in machine translation. In *Proceedings of LREC 2012*, pages 1208–1211, Istanbul.

Korkontzelos, Ioannis. 2011. Unsupervised Learning of Multiword Expressions. Ph.D. thesis, University of York, York.

Korkontzelos, Ioannis and Suresh Manandhar. 2010. Can recognising multiword expressions improve shallow parsing? In *Proceedings of NAACL/HLT 2010*, pages 636–644, Los Angeles, CA.

Krenn, Brigitte. 2008. Description of evaluation resource – German PP-verb data. In *Proceedings of the LREC 2008 Workshop on MWEs*, pages 7–10, Marrakech.

Krstev, Cvetana, Ivan Obradovic, Ranka Stankovic, and Dusko Vitas. 2013. An approach to efficient processing of multi-word units. In Przepiórkowski, A., Piasecki M., Jassem K., Fuglewicz P., editors, *Computational Linguistics - Applications*. Springer, pages 109–129.

Lambert, Patrik and Rafael Banchs. 2005. Data inferred multi-word expressions for statistical machine translation. In *Proceedings of Machine Translation Summit X*, pages 396–403, Phuket.

Lapata, Mirella and Alex Lascarides. 2003. Detecting novel compounds: The role of distributional evidence. In *Proceedings of EACL 2003*, pages 235–242, Budapest.

Laporte, Elena-Mirabela. 2014. *La traduction automatique statistique factorisée: une application à la paire de langues Français-Roumain*. Ph.D. thesis, University of Strasbourg.

Le Roux, Joseph, Antoine Rozenknop, and Matthieu Constant. 2014. Syntactic parsing and compound recognition via dual decomposition: Application to French. In *Proceedings of COLING 2014*, pages 1875–1885, Dublin.

Le Roux, Joseph, Antoine Rozenknop, and Jennifer Foster. 2013. Combining PCFG-LA models with dual decomposition: A case study with function labels and binarization. In *Proceedings of EMNLP 2013*, pages 1158–1169, Seattle, WA.

Legrand, Joël and Ronan Collobert. 2016. Phrase representations for multiword expressions. In *Proceedings of the ACL 2016 Workshop on MWEs*, pages 67–71, Berlin.

de Lhoneux, Miryam. 2015. CCG parsing and multiword expressions. *CoRR*, abs/1505.04420.

Losnegaard, Gyri Smørdal, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. 2016. Parseme survey on MWE resources. In *Proceedings of LREC 2016*, pages 2299–2306 Portoroz.

Mamede, Nuno J., Jorge Baptista, Cláudio Diniz, and Vera Cabarrão. 2012. STRING: An hybrid statistical and rule-based natural language processing chain for Portuguese. In *Proceedings of PROPOR 2012 Demonstrations*, Coimbra.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330.

Martens, Scott and Vincent Vandeghinste. 2010. An efficient, generic approach to extracting multi-word expressions from dependency trees. In *Proceedings of the COLING 2010 Workshop on MWEs*, pages 85–88, Beijing.

Martins, André F. T., Noah A. Smith, Eric P. Xing, Pedro M. Q. Aguiar, and Mário A. T. Figueiredo. 2010. Turbo parsers: Dependency parsing by approximate variational inference. In *Proceedings of EMNLP 2010*, pages 34–44, Cambridge, MA.

McCarthy, Diana, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 Workshop on MWEs*, pages 73–80, Sapporo.

McCarthy, Diana, Sriram Venkatapathy, and Aravind Joshi. 2007. Detecting compositionality of verb-object combinations using selectional preferences. In *Proceedings of EMNLP/CoNLL 2007*, pages 369–379, Prague.

McDonald, Ryan, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of HLT/EMNLP 2005*, pages 523–530, Vancouver.

McKeown, Kathleen R. and Dragomir R. Radev. 1999. Collocations. In H. Moisl, R. Dale, H. Somers, editors, *A Handbook of Natural Language Processing*. Marcel Dekker, New York, NY, chapter 15, pages 507–523.

Melamed, I. Dan. 1997. Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of EMNLP 1997*, pages 97–108, Providence, RI.

Mel'čuk, Igor, Nadia Arbatchewsky-Jumarie, André Clas, Suzanne Mantha, and Alain Polguère. 1999. *Dictionnaire explicatif et combinatoire du français contemporain.*

Recherches lexico-sémantiques IV. Les presses de l'Université de Montréal, Montréal, QC.
Mirroshandel, Seyed Abolghasem, Alexis Nasr, and Joseph Le Roux. 2012. Semi-supervised dependency parsing using lexical affinities. In *Proceedings of ACL 2012*, pages 777–785. Jeju Island.
Mititelu, V. Barbu and Elena Irimia. 2015. Description of the Romanian syntax within universal dependency project. In *Proceedings of Linguistic Resources and Tools for Processing the Romanian Language*, page 185, Iasi.
Monti, Johanna, Anabela Barreiro, Annibale Elia, Federica Marano, and Antonella Napoli. 2011. Taking on new challenges in multi-word unit processing for machine translation. In *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 11–19, Barcelona.
Monti, Johanna, Annibale Elia, Alberto Postiglione, Mario Monteleone, and Federica Marano. 2012. In search of knowledge: Text mining dedicated to technical translation. In *Proceedings of ASLIB 2011 Translating and the Computer Conference*, London.
Monti, Johanna, Federico Sangati, and Mihael Arcan. 2015. TED-MWE: A bilingual parallel corpus with MWE annotation. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, pages 193–197, Trento.
Morin, Emmanuel and Béatrice Daille. 2010. Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation*, 44(1-2):79–95.
Na, Hwidong, Jin-Ji Li, Yeha Lee, and Jong-Hyeok Lee. 2010. A synchronous context free grammar using dependency sequence for syntax-based statistical machine translation. In *Proceedings of AMTA 2010*, Denver, CO.
Nagy T., István and Veronika Vincze. 2014. VPCTagger: Detecting verb-particle constructions with syntax-based methods. In *Proceedings of the EACL 2014 Workshop on MWEs*, pages 17–25, Gothenburg.
Nasr, Alexis, Carlos Ramisch, José Deulofeu, and André Valli. 2015. Joint dependency parsing and multiword expression tokenisation. In *Proceedings of ACL 2015*, pages 1116–1126, Beijing.
Navigli, Roberto and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.
Nilsson, Jens, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of EMNLP/CoNLL 2007 CoNLL Shared Tasks Session*, pages 915–932, Prague.
Nivre, Joakim. 2014. Transition-based parsing with multiword expressions. In *Second PARSEME meeting*, Athens.
Nivre, Joakim, Yoav Goldberg, and Ryan McDonald. 2014. Constrained arc-eager dependency parsing. *Computational Linguistics*, 40(2):249–527.
Nivre, Joakim, Johan Hall, and Jens Nilsson. 2004. Memory-based dependency parsing. *Proceedings of CoNLL 2004*, pages 49–56, Boston, MA.
Nivre, Joakim and Jens Nilsson. 2004. Multiword units in syntactic parsing. In *Proceedings of Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA)*, pages 39–46.
Och, Franz Josef and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of ACL 2000*, pages 440–447, Hong Kong.
Oepen, Stephan, Helge Dyvik, Jan Tore Lønning, Erik Velldal, Dorothee Beermann, John Carroll, Dan Flickinger, Lars Hellan, Janne Bondi Johannessen, and Paul Meurer. 2004. Som å kapp-ete med trollet? Towards MRS-based Norwegian-English machine translation. In *Proceedings of TMI 2004*, pages 11–20, Baltimore, MD.
Oflazer, Kemal, Özlem Çetinoğlu, and Bilge Say. 2004. Integrating morphology with multi-word expression processing in Turkish. In *Proceedings of the ACL 2004 Workshop on MWEs*, pages 64–71, Barcelona.
Okita, Tsuyoshi. 2012. *Word Alignment and Smoothing Methods in Statistical Machine Translation: Noise, Prior Knowledge and Overfitting*. Ph.D. thesis, University of York.
Okita, Tsuyoshi and Andy Way. 2011. MWE-sensitive word alignment in factored translation model. In *Proceedings of MTML*, pages 16–17, Haifa.
Pal, Santanu, Sudip Naskar, and Sivaji Bandyopadhyay. 2013. A hybrid word alignment model for phrase-based statistical machine translation. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 94–101, Sofia.
Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318, Philadelphia, PA.

Pearce, Darren. 2001. Synonymy in collocation extraction. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, pages 41–46, Pittsburgh, PA.
Pearce, Darren. 2002. A comparative evaluation of collocation extraction techniques. In *Proceedings of LREC 2002*, pages 1530–1536, Canary Islands.
Pecina, Pavel. 2008. *Lexical Association Measures: Collocation Extraction*. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University in Prague, Prague.
Pecina, Pavel. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1-2):137–158.
Pedersen, Ted. 1996. Fishing for exactness. In *Proceedings of the South - Central SAS Users Group Conference (SCSUG 1996)*, pages 188–200, Austin, TX.
Pei, Wenzhe, Tao Ge, and Baobao Chang. 2015. An effective neural network model for graph-based dependency parsing. In *Proceedings of ACL 2015*, pages 313–322, Beijing.
Pollard, Carl and Ivan Sag. 1994. *Head-driven Phrase Structure Grammar*. Studies in Contemporary Linguistics, University of Chicago Press.
Pretkalnia, Lauma and Laura Rituma. 2012. Syntactic issues identified developing the Latvian treebank. In A. Tavast, K. Muischnek, and Koit, editors, *Human Language Technologies: The Baltic Perspective*, volume 247, IOS Press, Amsterdam. page 185.
Rácz, A., T. István Nagy, and Veronika Vincze. 2014. 4FX: Light verb constructions in a multilingual parallel corpus. In *Proceedings of LREC 2014*, pages 710–715, Reykjavik.
Ramisch, Carlos. 2015. *Multiword expressions acquisition: A generic and open framework*, volume XIV of *Theory and Applications of Natural Language Processing*. Springer.
Ramisch, Carlos, Laurent Besacier, and Alexander Kobzar. 2013. How hard is it to automatically translate phrasal verbs from English to French? In *MT Summit 2013 Workshop on Multi-word Units in Machine Translation and Translation Technology*, pages 53–61, Nice.
Ramisch, Carlos, Vitor De Araujo, and Aline Villavicencio. 2012. A broad evaluation of techniques for automatic acquisition of multiword expressions. In *Proceedings of the ACL 2012 Student Research Workshop*, pages 1–6, Jeju Island.
Ramisch, Carlos, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. 2008a. An evaluation of methods for the extraction of multiword expressions. In *Proceedings of the LREC 2008 Workshop on MWEs*, pages 50–53, Marrakech.
Ramisch, Carlos, Aline Villavicencio, and Christian Boitet. 2010. mwetoolkit: A framework for multiword expression identification. In *Proceedings of LREC 2010*, pages 662–669, La Valletta.
Ramisch, Carlos, Aline Villavicencio, and Valia Kordoni, editors. 2013. *ACM TSLP Special Issue on MWEs*, volume 10. ACM, New York, NY.
Ramisch, Carlos, Aline Villavicencio, Leonardo Moura, and Marco Idiart. 2008b. Picking them up and figuring them out: Verb-particle constructions, noise and idiomaticity. In *Proceedings of CoNLL 2008*, pages 49–56, Manchester.
Ramshaw, Lance and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94, Cambridge, MA.
Rayson, Paul, Scott Piao, Serge Sharoff, Stefan Evert, and Begoña Villada Moirón, editors. 2010. *LRE Special Issue on Multiword expression: Hard Going or Plain Sailing*, 44(1-2), Springer, Berlin.
Reddy, Siva, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of IJCNLP 2011*, pages 210–218, Chiang Mai.
Ren, Zhixiang, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the ACL 2009 Workshop on MWEs*, pages 47–54, Singapore.
Resnik, Philip. 1992. Probabilistic tree-adjoining grammar as a framework for statistical natural language processing. In *Proceedings of COLING 1992*, pages 418–424, Nantes.
Riedl, Martin and Chris Biemann. 2015. A single word is not enough: Ranking multiword expressions using distributional semantics. In *Proceedings of EMNLP 2015*, pages 2430–2440, Lisbon.
Riedl, Martin and Chris Biemann. 2016. Impact of MWE resources on multiword recognition. In *Proceedings of the ACL 2016 Workshop on MWEs*, pages 107–111, Berlin.
Riezler, Stefan, Tracy H. King, Ronald M. Kaplan, Richard Crouch, III John T. Maxwell, and Mark Johnson. 2002. Parsing

the Wall Street Journal using a lexical-functional grammar and discriminative estimation techniques. In *Proceedings of ACL 2002*, pages 271–278, Philadelphia, PA.
Rondon, Alexandre, Helena Caseli, and Carlos Ramisch. 2015. Never-ending multiword expressions learning. In *Proceedings of the ACL 2015 Workshop on MWEs*, pages 45–53, Denver, CO.
Rosén, Victoria, Koenraad De Smedt, Gyri Smørdal Losnegaard, Eduard Bejcek, Agata Savary, and Petya Osenova. 2016. MWEs in treebanks: From survey to guidelines. In *Proceedings of LREC 2016*, pages 2323–2330, Portoroz.
Rosén, Victoria, Gyri Smørdal Losnegaard, De Smedt Koenraad, Eduard Bejček, Agata Savary, Adam Przepiórkowski, Manfred Sailer, and Mitetelu Verginica. 2015. A survey of multiword expressions in treebanks. In *Proceedings of TLT14*, pages 179–193, Warsaw.
Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of CICLing 2002*, pages 1–15, Mexico City.
Sagot, Benoît and Pierre Boullier. 2005. From raw corpus to word lattices: Robust pre-parsing processing with SxPipe. *Archives of Control Sciences*, 15(4):653–662.
Salehi, Bahar, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of NAACL/HLT 2015*, pages 977–983, Denver, CO.
Salehi, Bahar, Nitika Mathur, Paul Cook, and Timothy Baldwin. 2015. The impact of multiword expression compositionality on machine translation evaluation. In *Proceedings of the ACL 2015 Workshop on MWEs*, page 54–59, Denver, CO.
Salton, Giancarlo, Robert Ross, and John Kelleher. 2014. Evaluation of a substitution method for idiom transformation in statistical machine translation. In *Proceedings of the EACL 2014 Workshop on MWEs*, pages 38–42, Gothenburg.
Sangati, Federico and Andreas van Cranenburgh. 2015. Multiword expression identification with recurring tree fragments and association measures. In *Proceedings of the ACL 2015 Workshop on MWEs*, pages 10–18, Denver, CO.
Savary, Agata. 2009. Multiflex: A multilingual finite-state tool for multi-word units. In *Proceedings of CIAA 2009*, pages 237–240, Sydney.
Savary, Agata and Adam Przepiorkowski. 2013. PARSEME: Parsing and Multi-Word Expressions. Towards linguistic precision and computational efficiency in Natural Language Processing. COST Action no. IC1207.
Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of EACL 2017 Workshop on MWEs*, pages 31–47, Valencia.
Savary, Agata, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartin, Jakub Waszczuk, Matthieu Constant, Petya Osenova, and Federico Sangati. 2015. PARSEME – Parsing and multiword expressions within a European multilingual network. In *Proceedings of LTC 2015*, Poznań.
Schneider, Gerold. 2012. Using semantic resources to improve a syntactic dependency parser. In *Proceedings of SEM-II workshop at LREC 2012*, pages 67–76, Istanbul.
Schneider, Nathan, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014a. Discriminative lexical semantic segmentation with gaps: running the MWE gamut. *TACL*, 2193–206.
Schneider, Nathan, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. Semeval-2016 task 10: Detecting minimal semantic units and their meanings (dimsum). In *Proceedings of SemEval 2016*, pages 546–559, San Diego, CA.
Schneider, Nathan, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014b. Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of LREC 2014*, pages 455–461, Reykjavik.
Schottmüller, Nina and Joakim Nivre. 2014. Issues in translating verb-particle constructions from German to English. In *Proceedings of the EACL 2014 Workshop on MWEs*, pages 124–131, Gothenburg.
Seddah, Djamé, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann,

Wolfgang Maier, Joakim Nivre, Adam Przepiorkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clérgerie. 2013. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of SPRML 2013*, pages 146–182, Seattle, WA.

Segura, J., and V. Prince. 2011. Using alignment to detect associated multiword expressions in bilingual corpora. In *Proceedings from Tralogy I*, Paris. `http://lodel.irevues.inist.fr/tralogy/index.php?id=144`.

Sekine, Satoshi and Michael Collins. 1997. EVALB bracket scoring program. `http://nlp.cs.nyu.edu/evalb/`.

Sennrich, Rico and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 83–91, Berlin.

Seretan, Violeta. 2011. *Syntax-Based Collocation Extraction*, Text, Speech and Language Technology, Springer.

Shigeto, Yutaro, Ai Azuma, Sorami Hisamoto, Shuhei Kondo, Tomoya Kouse, Keisuke Sakaguchi, Akifumi Yoshimoto, Frances Yung, and Yuji Matsumoto. 2013. Construction of English MWE dictionary and its application to POS tagging. In *Proceedings of the NAACL/HLT Workshop on MWEs*, pages 139–144, Atlanta, GA.

Silberztein, Max. 1997. The lexical analysis of natural languages. In E. Roche, Y. Schabes, editors, *Finite-State Language Processing*, MIT Press, pages 175–203.

da Silva, Joaquim Ferreira, Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. 1999. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In *Proceedings of the 9th Portuguese Conference on Artificial Intelligence*, pages 113–132, London.

Sinha, R. Mahesh K. 2009. Mining complex predicates in Hindi using a parallel Hindi-English corpus. In *Proceedings of the ACL 2009 Workshop on MWEs*, pages 40–46, Singapore.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Somers, Harold. 1999. Review article: Example-based machine translation. *Machine Translation*, 14(2):113–157.

Sporleder, Caroline and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of EACL 2009*, pages 754–762, Athens.

Steedman, Mark. 1987. Combinatory grammars and parasitic gaps. *Natural Language & Linguistic Theory*, 5(3):403–439.

Steinberger, Ralf, Bruno Pouliquen, Alexandrov Kabadjov, Mijail, and Erik Van der Goot. 2013. JRC-names: A freely available, highly multilingual named entity resource. *CoRR*, abs/1309.6162.

Stevenson, Suzanne, Afsaneh Fazly, and Ryan North. 2004. Statistical measures of the semi-productivity of light verb constructions. In *Proceedings of the ACL 2004 Workshop on MWEs*, pages 1–8, Barcelona.

Stymne, Sara, Nicola Cancedda, and Lars Ahrenberg. 2013. Generation of compound words in statistical machine translation into compounding languages. *Computational Linguistics*, 39(4):1067–1108.

Tan, Liling and Santanu Pal. 2014. Manawi: Using multi-word expressions and named entities to improve machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 201–206, Baltimore, MD.

Tsvetkov, Yulia and Shuly Wintner. 2014. Identification of multiword expressions by combining multiple linguistic information sources. *Computational Linguistics*, 40(2):449–468.

Tu, Yuancheng. 2012. English Complex Verb Constructions: Identification and Inference. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Tutin, Agnes, Emmanuelle Esperana-Rodier, Manolo Iborra, and Justine Reverdy. 2015. Annotation of multiword expressions in French. In *Proceedings of Europhras 2015 (Europhras15)*, pages 60–67, Malaga.

Uchiyama, Kiyoko, Timothy Baldwin, and Shun Ishizaki. 2005. Disambiguating Japanese compound verbs. *CSL Special Issue on MWEs*, 19(4):497–512.

Ullman, Edvin and Joakim Nivre. 2014. Paraphrasing Swedish compound nouns in machine translation. In *Proceedings of the EACL 2014 Workshop on MWEs*, pages 579–587, Gothenburg.

Urieli, Assaf. 2013. Robust French Syntax Analysis: Reconciling Statistical Methods and Linguistic Knowledge in the Talismane Toolkit. Ph.D. thesis, University of Toulouse II le Mirail.

Villavicencio, Aline, Francis Bond, Anna Korhonen, and Diana McCarthy, editors. 2005. *CSL Special Issue on MWEs*, volume 19.

Villavicencio, Aline, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of EMNLP/CoNLL 2007*, pages 1034–1043, Prague.

Villemonte De La Clergerie, Éric. 2013. Improving a symbolic parser through partially supervised learning. In *Proceedings of IWPT 2013*, Nara.

Vincze, Veronica, István Nagy, and Gábor Berend. 2011. Multiword expressions and named entities in the Wiki50 corpus. In *Proceedings of RANLP 2011*, pages 289–295, Hissar.

Vincze, Veronika. 2012. Light verb constructions in the SzegedParalellFX English–Hungarian parallel corpus. In *Proceedings of LREC 2012*, pages 2381–2388, Istanbul.

Vincze, Veronika, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian dependency treebank. In *Proceedings of LREC 2010*, volume 10, pages 1855–1862, Malta.

Vincze, Veronika, János Zsibrita, and Istvàn Nagy T. 2013. Dependency parsing for identifying Hungarian light verb constructions. In *Proceedings of IJCNLP 2013*, pages 207–215, Nagoya.

Volk, Martin. 2001. Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In *Proceedings of Corpus Linguistics 2001*, pages 601–606, Lancaster.

Schulte im Walde, Sabine, Stefan Müller, and Stefan Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Proceedings of *SEM 2013*, pages 255–265, Atlanta, GA.

Wehrli, Eric. 2014. The relevance of collocations for parsing. In *Proceedings of the EACL 2014 Workshop on MWEs*, pages 26–32, Gothenburg.

Wehrli, Eric, Violeta Seretan, and Luka Nerima. 2010. Sentence analysis and collocation identification. In *Proceedings of the COLING 2010 Workshop on MWEs*, pages 28–36, Beijing.

Wehrli, Eric, Violeta Seretan, Luka Nerima, and L. Russo. 2009. Collocations in a rule-based MT System: A case study evaluation of their translation adequacy. In *Proceedings of EAMT 2009*, pages 128–135, Barcelona.

Wei, Wei and Bo Xu. 2011. Effective use of discontinuous phrases for hierarchical phrase-based translation. In *MT Summit XIII*, pages 397–404, Xiamen.

Weller, Marion, Fabienne Cap, Stefan Muller, Sabine Schulte im Walde, and Alexander Fraser. 2014. Distinguishing degrees of compositionality in compound splitting for statistical machine translation. In *Proceedings of the COLING Workshop on Computational Approaches to Compound Analysis*, pages 81–90, Dublin.

Weller, Marion and Ulrich Heid. 2010. Extraction of German multiword expressions from parsed corpora using context features. In *Proceedings of LREC 2010*, pages 3195–3201, Valletta.

Wray, Alison. 2002. *Formulaic Language and the Lexicon*, Cambridge University Press, Cambridge.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144:1–23.

XTAG. 2001, A lexicalized tree adjoining grammar for English. Technical Report IRCS-01-03, IRCS, University of Pennsylvania.

Yazdani, Majid, Meghdad Farahmand, and James Henderson. 2015. Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of EMNLP 2015*, pages 1733–1742, Lisbon.

Zarrieß Sina and Jonas Kuhn. 2009. Exploiting translational correspondences for pattern-independent MWE identification. In *Proceedings of the ACL 2009 Workshop on MWEs*, pages 23–30, Singapore.