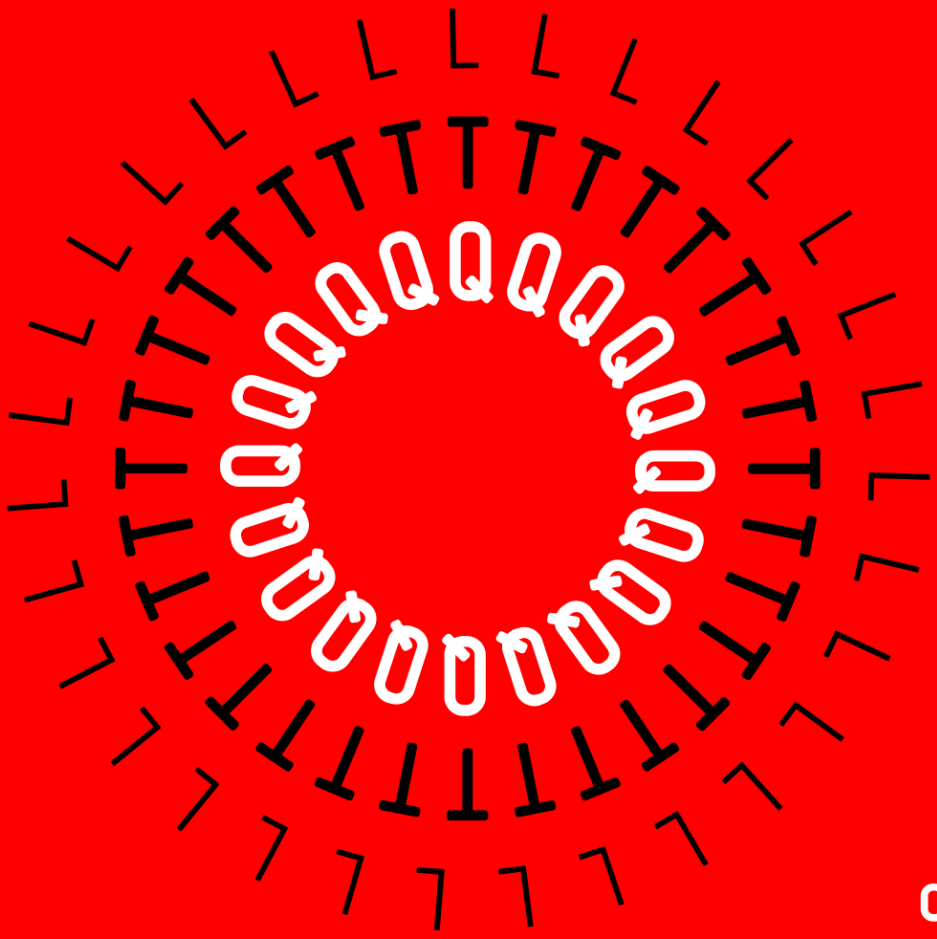


qtleap
quality translation by deep language engineering approaches

Report on the State of the Art Concerning Multiword Expressions

DELIVERABLE D4.3

VERSION 7.0 | 2015-06-07

QTLeap

Machine translation is a computational procedure that seeks to provide the translation of utterances from one language into another language.

Research and development around this grand challenge is bringing this technology to a level of maturity that already supports useful practical solutions. It permits to get at least the gist of the utterances being translated, and even to get pretty good results for some language pairs in some focused discourse domains, helping to reduce costs and to improve productivity in international businesses.

There is nevertheless still a way to go for this technology to attain a level of maturity that permits the delivery of quality translation across the board.

The goal of the QTLeap project is to research on and deliver an articulated methodology for machine translation that explores deep language engineering approaches in view of breaking the way to translations of higher quality.

The deeper the processing of utterances the less language-specific differences remain between the representation of the meaning of a given utterance and the meaning representation of its translation. Further chances of success can thus be explored by machine translation systems that are based on deeper semantic engineering approaches.

Deep language processing has its stepping-stone in linguistically principled methods and generalizations. It has been evolving towards supporting realistic applications, namely by embedding more data based solutions, and by exploring new types of datasets recently developed, such as parallel DeepBanks.

This progress is further supported by recent advances in terms of lexical processing. These advances have been made possible by enhanced techniques for referential and conceptual ambiguity resolution, and supported also by new types of datasets recently developed as linked open data.

The project QTLeap explores novel ways for attaining machine translation of higher quality that are opened by a new generation of increasingly sophisticated semantic datasets and by recent advances in deep language processing.

www.qtleap.eu

Funded by

QTLeap is funded by the 7th Framework Programme of the European Commission.

Supported by

And supported by the participating institutions:

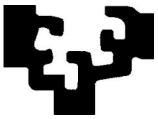
Faculty of Sciences, University of Lisbon

German Research Centre for Artificial Intelligence

Charles University in Prague

Bulgarian Academy of Sciences

Humboldt University of Berlin



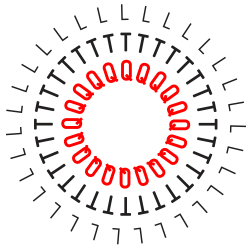
University of Groningen

Higher Functions, Lda

Revision history

Version	Date	Authors	Organisation	Description
1.0	Jan 14, 2014	Kostadin Cholakov, Valia Kordoni, Markus Egg	UBER	First draft
1.5	Jan 17, 2014	Kostadin Cholakov, Valia Kordoni, Markus Egg, Kepa Sarasola	UBER, UPV-EHU	Feedback from UPV-EHU integrated
2.0	Jan 22, 2014	Kostadin Cholakov, Valia Kordoni, Markus Egg, Kepa Sarasola, Petya Osenova, Gertjan van Noord	UBER, UPV-EHU, UG, IICT-BAS	Feedback from UG and IICT-BAS integrated
3.0	Jan 27, 2014	Kostadin Cholakov, Valia Kordoni, Markus Egg, Kepa Sarasola, Petya Osenova, Gertjan van Noord	UBER, UPV-EHU, UG, IICT-BAS	Feedback from internal review integrated
3.5	Feb 3, 2014	Kostadin Cholakov, Valia Kordoni, Markus Egg, Kepa Sarasola, Petya Osenova, Gertjan van Noord, António Branco	UBER, UPV-EHU, UG, IICT-BAS, FCUL	Feedback from FCUL integrated
4.0	Feb 6, 2014	Kostadin Cholakov, Valia Kordoni, Markus Egg, Kepa Sarasola, Petya Osenova, Gertjan van Noord, António Branco	UBER, UPV-EHU, UG, IICT-BAS, FCUL, DFKI	Minor comments from DFKI addressed
5.0	Apr 29, 2014	Kostadin Cholakov, Valia Kordoni, Markus Egg, Kepa Sarasola, Petya Osenova, Gertjan van Noord, António Branco	UBER, UPV-EHU, UG, IICT-BAS, FCUL, DFKI	Modified in accordance with the recommendations made in the technical audit
6.0	Oct 8, 2014	Kostadin Cholakov	UBER	Added chapter "Expected Benefits for the Real User Scenario"
6.0	Oct 10, 2014	Kostadin Cholakov	UBER	All relevant partners agreed with the changes made in this version and no further input was provided
7.0	Jun 7, 2015	Kostadin Cholakov	UBER	Changes to address the comments in the first year review

Statement of originality
This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.



REPORT ON THE STATE OF THE ART CONCERNING MULTIWORD EXPRESSIONS

DOCUMENT QTLEAP-2015-D4.3

EC FP7 PROJECT #610516

DELIVERABLE D4.3

completion

FINAL

status

RESUBMITTED

dissemination level

PUBLIC

responsible

KIRIL SIMOV (WP4 COORDINATOR)

reviewer

ENEKO AGIRRE

contributing partners

FCUL, DFKI, IICT-BAS, UBER, UPV-EHU, UG

authors

KOSTADIN CHOLAKOV, VALIA KORDONI, MARKUS EGG,

KEPA SARASOLA, PETYA OSENOVA, GERTJAN VAN NOORD,

FRANCISCO COSTA, JOÃO SILVA, ALJOSCHA BURCHARDT

Contents

1 Introduction

Multiword expressions (MWEs) are units which consist of several lexemes but whose meaning is not derivable, or is only partly derivable from the semantics of their constituents. Some examples include idiomatic expressions such *take advantage of*, nominal compounds such as *traffic light*, and phrasal verbs such as *give up*. MWEs play an important role in every day language communication. Jackendoff (1997) estimates that the amount of MWEs in a speaker's lexicon is nearly the same as the amount of single words.

The high frequency of usage and the idiosyncratic semantic and syntactic properties of these constructions indicate the need for their special handling in Natural Language Processing (NLP). From the perspective of semantics, MWEs need to be treated as units, because their meaning spans over word boundaries. From the perspective of syntax, however, these expressions are often hard to identify, because of their resemblance to ordinary verb or noun phrases. The MWE *kick the bucket*, for instance, which on the syntax level is just an ordinary verb phrase, can receive a very different semantic interpretation than the intended one, if not treated as a unit (Sag et al., 2002).

The idiosyncrasy of MWEs, together with their frequent usage in every day language call for special treatment of those expressions in NLP. For example, deep linguistic grammars often treat MWEs as words-with-spaces, i.e., an MWE is added as a single lexical entry to the lexicon instead of falling back on rules in the grammar which would construct this MWE compositionally. This approach is very problematic, though. Consider, for example, light verb constructions which often form families: *take a walk*, *take a hike*, *take a flight*, etc. Listing such expressions individually leads to severe loss of generality and lack of prediction on the part of the grammar. In the context of the QTLeap project, robust deep linguistic processing clearly needs more sophisticated ways for interpreting MWEs.

In the context of machine translation (MT), MWEs pose an additional challenge, namely asymmetry. It is often the case that an MWE in the source language does not have an exact translation equivalent in the target language. Therefore, in the context of the QTLeap project, it is not only important to have better techniques for handling MWEs during monolingual parsing but it is also vital to improve the transfer of MWEs in the MT system. Better transfer techniques for MWEs could improve translation quality significantly.

Finally, we would like to note that in the context of this deliverable, we limit the notion of MWEs to the lexical level, i.e. to multi-word lexical units which need dedicated lexical processing techniques within the QTLeap project.

2 Linguistic Properties of Multiword Expressions

Before presenting the various approaches for treating MWEs, it is important to understand their main linguistic properties.

2.1 Definition

Research on MWEs is a long-standing linguistic enterprise, hence, many definitions about what is an MWE have been proposed. Given the limitation to lexical processing stated above, we will adopt the definition provided in Sag et al. (2002):

(1) Multiword expressions (MWEs) are lexical items that: (a) can be decomposed into

multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity.

Note that for languages such as German this definition needs to be relaxed and allow for single-word MWEs in order to handle the high productivity of compound nouns such as *Kontaktlinse* "contact lens" (the concatenation of *Kontakt* "contact" and *Linse* "lens") in which there is no white space delimitation. Also, note that though the definition speaks of *lexical units*, this does not necessarily mean *constituents*. If dependency parsing is employed, an MWE might be a catena in a dependency tree, such as in the case of *take advantage of*.

2.2 Idiomaticity

2.2.1 Lexical Idiomaticity

Lexical idiomaticity occurs when one or more parts of an MWE are not part of the conventional lexicon. For example, in *ad hoc*, neither of the components (*ad* and *hoc*) are standalone English words. Consequently, lexical idiomaticity inevitably results in syntactic and semantic idiomaticity because there is no lexical knowledge associated directly with the parts from which to predict the behaviour of the MWE.

2.2.2 Syntactic Idiomaticity

Syntactic idiomaticity occurs when the syntax of the MWE is not derived directly from that of its components (Sag et al., 2002). For example, *by and large*, is syntactically idiomatic in that it is adverbial in nature, but made up of the anomalous coordination of a preposition (*by*) and an adjective (*large*). This is opposite to the case of *take a break*, for instance, which is a normal verb-object combination derived from a transitive verb and a countable noun.

2.2.3 Semantic Idiomaticity

Semantic idiomaticity is the property of the meaning of an MWE not being explicitly derivable from its parts (Sag et al., 2002). For example, the meaning of *kick the bucket* (to die) cannot be predicted by either *kick* or *bucket*. On the other hand, *to and fro* is not semantically marked as its semantics is fully predictable from its parts.

Note that related to the issue of semantic idiomaticity is the notion of figurative language use, i.e. the meaning of an MWE can be metaphorical (e.g., *take the bull by the horns*), hyperbolic (e.g., *not worth the paper it's printed on*), or metonymic in nature (e.g., *lend a hand*) (Fillmore et al., 1988; Nunberg et al., 1994).

2.2.4 Pragmatic Idiomaticity

Pragmatic idiomaticity is the condition of an MWE being associated with a fixed set of situations or a particular context (Jackendoff, 1997; Sag et al., 2002). Examples of pragmatically marked MWEs include *good morning* and *all aboard*. The former is a greeting associated specifically with mornings and the latter is a command associated with the specific situation of a train station or dock, and the imminent departure of a train or ship.

2.2.5 Statistical Idiomaticity

We talk of statistical idiomaticity when a particular combination of words occurs with markedly high frequency, relative to the component words or alternative phrasings of the same expression. Cruse (1986) illustrates statistical idiomaticity in an experiment which examines a cluster of near-synonym adjectives and their affinity to pre-define certain nouns. One example is *impeccable credentials* which occurs much more frequently than *spotless credentials*. Another case of statistical idiomaticity are binomials such as *black and white* (as in *black and white television*) where the reverse noun ordering does not preserve the lexicalised semantics of the word combination (?*white and black television*) (Benor and Levy, 2006). Note that statistically idiomatic expressions are not semantically idiomatic. Thus, those pose a bigger challenge for natural language generation.

2.3 Multiword Expressions and Compositionality

Another important characteristic of MWEs is their compositionality. Compositionality is the degree to which the features of the parts of a given MWE combine to predict the features of the whole MWE.

Generally, compositionality is considered in a semantic context (Nunberg et al., 1994) but we can equally talk about lexical, syntactic, or pragmatic compositionality. MWEs such as *spill the beans*, for example, can be analysed as being made up of *spill* in a "reveal" sense and *the beans* in a "secret(s)" sense, resulting in the overall compositional reading of "reveal the secret(s)". On the other hand, no such analysis is possible with *kick the bucket*. Based on these observations, researchers (Sag et al., 2002) distinguish between three classes of MWEs: non-decomposable MWEs (e.g., *shoot the breeze*), idiosyncratically decomposable MWEs (e.g., *let the cat out of the bag*), and decomposable MWEs (e.g., *kindle excitement*).

The syntactic flexibility of an MWE can generally be explained in terms of its decomposability. For example, due to their opaque semantics, non-decomposable MWEs are not subject to syntactic variability such as internal modification or passivisation. The only types of modification such MWEs allow is inflection (e.g., *kicked the bucket*) and variation in reflexive forms (e.g., *wet oneself*).

2.4 Classification of Multiword Expressions

Based on the syntactic and semantic properties of MWEs, Sag et al. (2002) propose a commonly used high-level classification of MWEs. This classification is presented in Figure 1.

The class of institutionalised phrases corresponds to MWEs which are exclusively statistically idiomatic, for example *salt and pepper* and *many thanks*. Lexicalised phrases, on the other hand, are explicitly encoded in the lexicon and represent MWEs with lexical, syntactic, semantic or pragmatic idiomaticity.

Fixed expressions are fixed strings that undergo neither morphosyntactic variation nor internal modification, often due to fossilisation of what was once a compositional phrase. For example, *by and large* is not morphosyntactically modifiable (e.g., **by and larger*) or internally modifiable (e.g. *by and very large*). Non-modifiable prepositional phrases without a preceding determiner such as *on air* are also fixed expressions.

Semi-fixed expressions are MWEs that have hard restrictions on word order and composition, but undergo some degree of lexical variation (e.g., *kick/kicks/kicked the*

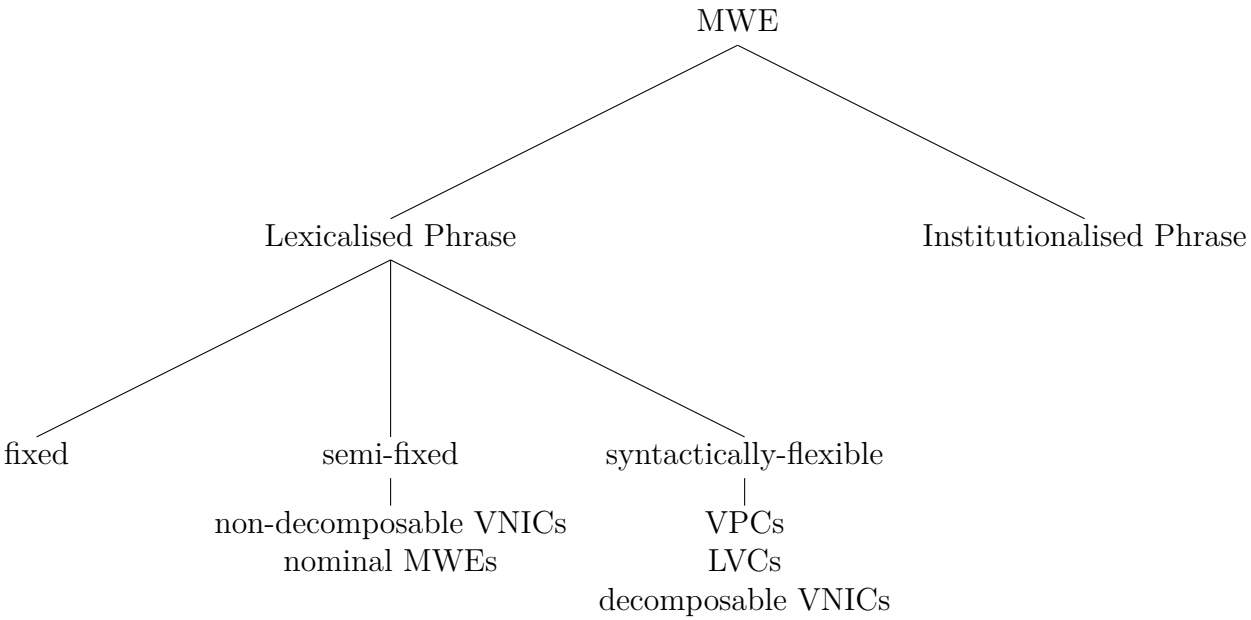


Figure 1: A classification for MWEs (Sag et al., 2002).

bucket vs. **the bucket was kicked*). Non-decomposable verb-noun idiomatic constructions (VNICs) such as *shoot the breeze* and nominal MWEs such as *attorney general* are also classified as semi-fixed expressions.

Syntactically flexible expressions are MWEs which undergo syntactic variation, such as verb particle constructions (VPCs), light verb constructions (LVCs) and decomposable VNICs. The nature of the flexibility varies significantly across construction types. VPCs, for example, are syntactically flexible with respect to the word order of the particle and the object in transitive usages: *hand in the paper* vs. *hand the paper in*. LVCs are even more flexible, allowing for example for internal modification: *take a walk* and *take a nice walk*.

Next, we describe the main research trends in NLP regarding the identification and computational processing of MWEs.

3 Identification and Extraction of Multiword Expressions

In order for NLP applications to handle properly MWEs, those expressions should be first identified. Various automated techniques have been proposed to either identify MWEs in running text or to extract them from predefined corpora. Below, we describe the main trends set by those techniques. The majority of those techniques treat MWEs as single units, with more or less fixed and stable structure. In order to handle MWEs with fully flexible structure, more sophisticated approaches are needed. However, as we will show in Section 7, the two major types of MWEs encountered in the technical domain from which the QTLeap data come are phrasal verbs and noun compounds/collocations. Those have relatively stable structure and which makes them easier to handle.

3.1 Extraction

MWE extraction is a task in which the MWEs attested in a predetermined corpus are extracted out into a lexicon. Thus the motivation for performing this task is mainly lexicon development, i.e. adding new types of MWEs to the lexicon. Although the task is closely related to that of MWE identification, extraction is different in the sense that it operates on a type level. For example, identification determines whether *give up* is an MWE in a particular sentence in the corpus whereas extraction determines whether the MWE *give up* occurs in the corpus or not.

There has been a strong focus on the development of general-purpose techniques for MWE extraction. The majority of those techniques are inspired by the idea that a word is 'characterised by the company it keeps' (Firth, 1957), i.e. MWEs can be found by looking at the frequency with which a combination of words occurs. The hypothesis is that the more frequently some words occur together, the more likely it is that they form an MWE. The analysis of co-occurrence is carried out by using various types of association measures such as t-test or pointwise mutual information (Church and Hanks, 1990), often in comparison with the frequencies of the component words of a candidate MWE. Association measures provide a score for each word combination, which forms the basis of a ranking of MWE candidates. Final extraction, therefore, consists of determining an appropriate cut-off in the ranking, although evaluation is often carried out over the full ranking. Pecina (2008) compares 84 association measures, finding out that some of those are rank equivalent. Ramisch et al. (2008) experiments with combinations of measures determining that some of those perform better than any single measure.

A second approach to MWE extraction, targeted specifically at semantically and statistically idiomatic MWEs, is to extend the general association measure approach to include substitution (Lin, 1999; Schone and Jurafsky, 2001; Pearce, 2001). For example, in assessing the idiomaticity of *red tape*, explicit comparison is made with lexically-related candidates generated by component word substitution, such as *yellow tape* or *red strip*. Common approaches for determining substitution candidates for a given component word are resources based on synonymity such as WordNet (Fellbaum, 1998) or distributional similarity.

However, the indiscriminate use of association measures is criticized by Dunning (1993). He argues that most measures assume a normal distribution for the words in a language but corpus evidence does not support this hypothesis. Therefore, he proposes a 2-gram measure called a likelihood ratio that estimates directly how much more likely a 2-gram is than one expected by chance. In addition to being theoretically sound, the scores calculated with this measure are also easily interpretable. Measures based on a likelihood ratio (e.g., the log-likelihood score) are employed in a fair number of current MWE extraction techniques.

Another statistical measure, permutation entropy, is presented in Zhang et al. (2006). It exploits the statistical idiomaticity of MWEs. The authors assume that if a given expression is just the result of the random occurrence of very frequent words, most probably the order of the words in this expression is not important. Therefore, they measure the frequency of occurrence of all permutations of that expression (e.g., *the likes of*: *of the likes*, *likes of the*, etc.) and calculate the sum of the entropies of all permutations. A high entropy means that permutations are more or less equally probable which the authors consider a clear indication of a random nature. This makes it highly unlikely that the expression in question is an idiom. On the other hand, if the entropy is close to 0, only a single configuration of the words is likely to be allowed, indicating that the expression is

probably an MWE.

One alternative approach is to use linguistic preprocessors such as POS taggers, parsers, chunkers, etc to first identify MWEs in a corpus and then feed the predictions of the various preprocessors into a statistical classifier which renders the final decision whether a given word combination is an MWE or not. Such techniques have been successfully applied to extract VPCs or determinerless preposition-noun combinations such as *on air* (Baldwin, 2005; van der Beek, 2005).

Another alternative approach follows the hypothesis that compositionality can be related to distributional similarity for MWE extraction. In this approach, the contexts of the MWE candidate are compared against the contexts of its corresponding components by means of different techniques. Gurrutxaga and Alegria (2012) propose similarity measures commonly used in Information Retrieval (VSM, LSA, Indri index, etc). Other authors such as de Medeiros et al. (2010) identify and extract MWEs in a multilingual context based on word alignment processes.

Finally, note that all extraction techniques are poorly equipped to handle less frequent MWEs since statistical measures for those are not that reliable. In classification, such MWEs also cause data sparseness problems. That is why, most extraction methods consider only word combinations which have a higher number of occurrences than some predetermined threshold.

3.2 Identification

Identification is the task of determining individual occurrences of MWEs in running text. A key challenge in this task is the ability to differentiate between MWEs and literal usages for word combinations. Consider the following example involving the phrase *make a face* (taken from Baldwin and Kim (2010)):

(2) a. Kim made a face at the policemen
b. Kim made a face in pottery class

in which the first sentence clearly contains an MWE. However, the same phrase is not an MWE in the second sentence. There, its lexemes must be interpreted literally.

Most of the techniques for identifying MWEs tend to be specific to a particular language or type(s) of MWEs. However, there have also been attempts to develop more general and language independent methods. The most obvious such method is to use a part-of-speech (POS) tagger, parser, chunker or some other language resource the output of which contains the lexical information necessary to identify MWEs. For example, a chunker or a phrase structure parser can be used to identify VPCs in English (McCarthy et al., 2003; Lapata and Lascarides, 2003; Kim and Baldwin, 2010). First, all POS tags or syntactic structures which mark a particle are identified and then, some heuristics are used (e.g., looking left of the particle within some fixed size word window) to discover the head of that particle. Such techniques, however, cannot handle the cases like the example presented above. Since both the MWE and the literal usage of the combination of its lexemes have the same surface syntactic representation, an MWE will be wrongly identified in the second sentence. On the other hand, deep parsers which have lexical entries for MWEs and disambiguate to the level of lexical items are able to make this distinction, via supertagging or full parsing (Baldwin et al., 2004; Blunsom, 2007). Such parsers, however, are dependent on the lexical coverage of their lexicons with regard to MWEs. If an MWE is not interpreted as such in the lexicon, the parser will most likely

fail to identify it during parsing.

Another general approach to MWE identification is to treat literal and MWE usages as different senses of a given word combination. This allows for the application of word sense disambiguation (WSD) techniques to the identification problem for MWEs. As with WSD research, both supervised (Patrick and Fletcher, 2005; Hashimoto and Kawahara, 2008) and unsupervised (Birke and Sarkar, 2006; Katz and Giesbrecht, 2006; Sporleder and Li, 2009) approaches have been applied to solve the problem. Although supervised techniques tend to be more accurate, those require a substantial number of annotated MWEs and literal combinations of words which are used as training data for statistical classifiers. The annotation of such instances is expensive and time-consuming and that is why unsupervised methods have predominantly been used.

Finally, a third approach, targeted particularly at semantically idiomatic MWEs, is to assume that MWEs occur in canonical forms or only in particular syntactic configurations, and do not undergo the same level of syntactic variation as literal usages. A strong example for such MWEs are non-decomposable VNICs, such as *kick the bucket* which is not internally modifiable and cannot occur in passive. Then, the challenge lies in determining the degree of the syntactic variability that can occur within such an MWE (Fazly et al., 2009).

4 Treating Multiword Expressions in Deep Grammars and Machine Translation Systems

Since the goal of QTLeap is to integrate deep linguistic processing in MT, we will focus on techniques handling MWEs within deep grammars and MT systems. However, a wide range of other NLP applications will also benefit if they provide for a special treatment of MWEs, e.g., topic modelling (Lau et al., 2013) or information retrieval (Acosta et al., 2011).

4.1 Representation

When developing techniques for the automated treatment of MWEs in deep linguistic processing, one must first decide how MWEs should be represented in the lexicon of the grammar in question as well as in deep treebanks annotated with that grammar since those are often used as training data. For example, contiguous and non-compositional MWEs such as *ad hoc* can often be added as words-with-spaces lexical entries. On the other hand, such an approach is inappropriate for non-contiguous expressions and compositional MWEs.

Further, in QTLeap, the representation of MWEs in parsed data should not cause additional problems for the translation system. For example, even if the grammar recognises *kick the bucket* as an MWE, it might still analyse *the bucket* as the direct object of *kick*. In this case, translation becomes harder because the analysis of the phrase does not indicate that this is an MWE. Therefore, the correct analysis should clearly indicate MWEs, for example by treating *kick the bucket* as a single node in the parse tree.

Various representation approaches have been proposed. The most intuitive one is to include MWEs as words-with-spaces in the lexicon (e.g., *kick_the_bucket*). As mentioned above, this approach will not work for non-contiguous MWEs. Further, some types of

MWEs such as LVCs (e.g. *give up*, *give in*, etc.) are too productive and listing them in the lexicon will curb the ability of the grammar for making linguistic generalisations.

Parsing systems which output dependency structures are better equipped to deal with such cases. If the grammar recognises an MWE, the words forming this expression are often connected into a single node in the dependency tree. Additionally, attribute-value grammar formalisms such as HPSG are often able to provide some sort of deeper semantic analysis. For example, Copestake et al. (2002) describe the handling of LVCs and English noun compounds in the ERG where the lexical entries for those types of MWEs include special semantic features which link the units of the MWE together on semantic level. Accordingly, in MT based on semantic transfer, this would be a proper way of handling MWEs. However, the automated creation of such lexical entries is not a trivial task.

Finally, apart from word order, there are also other linguistic phenomena which must be considered: morphology (*kick the bucket* vs *kicked the bucket*), alternations in the MWE (*hand in the paper* vs. *hand the paper in*). To summarise, two conditions must be fulfilled when deciding on the representation of MWEs. First, the structure of the parser output should not obstruct the marking of an MWE. Second, the representation should be general enough to handle morphological and lexical variations of the same MWE but also specific enough to recognise only the MWE it is designed for.

4.2 State of the Art MWEs Techniques for Treatment of MWEs

Techniques in parsing. Baldwin et al. (2004) and van Noord (2004) show that missing or wrong lexical entries often cause parsing failures when parsing with deep grammars. A significant number of those entries represent MWEs. In order to address this problem Zhang et al. (2006) employ the error mining technique presented in van Noord (2004) to detect semi-automatically MWE candidates in texts. Then, those candidates are validated using a combination of the World Wide Web as a corpus and the permutation entropy presented in the previous section. Finally, a statistical classifier assigns lexical entries to the remaining candidates. Note that this approach adds the MWEs as words-with-spaces in the lexicon. Despite the shortcomings of this type of lexicon extension which we mentioned earlier, Zhang et al. (2006) show that their approach improves the performance of the English Resource Grammar (ERG; Copestake and Flickinger (2000)) which is a deep grammar of English.

Villavicencio et al. (2007) improve on this approach by adding a new lexical entry only for what they refer to as the 'head of the MWE'. Once a proper lexical entry for the head of the MWE is added, the whole expression can be analysed in a compositional way, i.e. the rules of the grammar are used to combine the lexemes of the MWE. For example, the expression *foot the bill* will be correctly handled, if there is a transitive verb reading for the word *foot* in the lexicon. Various heuristics are employed to identify the head of the MWE and then, the classification setup described in Zhang et al. (2006) is used to assign lexical entries to the head words only. Experiments with the ERG grammar show that this approach creates linguistically more general lexical entries and it further improves the grammar performance on a test set containing MWEs.

Constant et al. (2013) investigate possible ways of incorporating MWEs into a parser, focusing on French contiguous MWEs. They perform MWE identification prior to parsing and use a grammar that includes MWE identification via specialized annotation schemes for compounds. These solutions bring improvements that are often reflected in parsing accuracy.

Techniques for MT transfer. Recent research has demonstrated that even the incorporation of simple treatment for MWEs in MT systems may improve translation quality. For example, Carpuat and Diab (2010) adopt two complementary strategies: a static strategy of single-tokenisation, that treats MWEs as words-with-spaces, and a dynamic strategy, that keeps a record of the number of MWEs in the source phrase. They find that both strategies result in improvement of translation quality, which suggests that statistical MT based on phrase alignment alone may not model all MWE information.

Improvements are made in Pal et al. (2010) who applies preprocessing steps like single-tokenisation along with prior alignment and transliteration for named entities and compound verbs. Morin and Daille (2010) obtains an improvement of 33% in the French-Japanese translation of MWEs with a morphologically-based compositional method for backing off when there is not enough data in a dictionary to translate an MWE. For example, *chronic fatigue syndrome* is decomposed as [*chronic fatigue*] [*syndrome*], [*chronic*] [*fatigue syndrome*] or [*chronic*] [*fatigue*] [*syndrome*].

Finally, Simova and Kordoni (2013) investigate the very interesting and important issue of asymmetry in MT. In the context of MWEs, asymmetry occurs when an MWE in the source language does not have an equivalent MWE in the target one. The work investigates the case of English phrasal verbs when translating to Bulgarian, a language without such verbs. In most cases, a phrasal verb in English is translated into a single Bulgarian verb which causes alignment issues for statistical MT systems. Simova and Kordoni (2013) use lexicon lookup and shallow parsing to identify phrasal verbs in the English data and then make modifications in a statistical MT system to account for those verbs. As a result, translation quality improves significantly.

5 The MWE Community

The previous sections described considerable amount of research on the theoretical aspects of MWEs as well as their treatment within NLP. In this section, we would like to emphasise on the fact that there is a thriving research community in the center of MWE-related research.

Researchers from several fields view MWEs as a key problem in current NLP technology and yet, there are still important and urgent open matters to be solved. Today the MWE research community is organized as a Section of SIGLEX. We would like to note that one of the authors of this deliverable, Valia Kordoni, is the president of this Section and as such she is also the representative of the MWE community to the SIGLEX board.

The first and most important place to exchange ideas on MWE research is the annual workshop on MWEs. It is a series of workshops that have been held since 2001 in conjunction with major computational linguistics conferences. The recent editions of the workshop show that there is a shift from research on identification- and extraction-methods work toward more application-oriented research. The evaluation of MWE processing techniques and multilingual aspects are also current issues in the field.

Most of the information concerning past editions of the MWE workshop series can be found at the MWE Section website.[1] The site also hosts a repository with several annotated datasets and a list of software capable of dealing with MWEs. The community also adopted a mailing list to which anyone can subscribe. In addition to the dedicated

[1] `http://multiword.sourceforge.net`

workshop series, the *SEM conference[2] features a track on MWEs and all main computational linguistics conferences such as COLING, ACL, and LREC regularly feature papers on MWEs.

As a complement to workshops and conferences, there have been 3 special issues published by leading journals in computational linguistics. The first such issue was the *Journal of Computer Speech and Language* (Villavicencio et al., 2005), the second was the *Journal of Language Resources and Evaluation* (Rayson et al., 2010), and the last one was the *ACM Transactions on Speech and Language Processing* (Ramisch et al., 2013). Special issues like these provide a broad overview and present the most relevant research results coming from different authors and research groups working on the subject.

6 Beyond the State of the Art

As outlined in the previous section, research in statistical MT has already turned its interest to MWEs. Specifically, phrase-based SMT systems which rely solely on mappings of the level of word sequences may produce inadequate translations in cases where complex linguistic phenomena occur, among them some types of MWEs. For accurate, precise, and natural high-quality translation results, statistical MT has acknowledged that it is important to incorporate an adequate treatment of MWEs, whose interpretation poses a challenge given their idiosyncratic, flexible, and heterogeneous nature. The techniques of Carpuat and Diab (2010) and Pal et al. (2010) clearly demonstrate that even the incorporation of a simple treatment for MWEs in MT systems improves translation quality.

Characteristic of this trend of research on exploitation of MWEs for improvement of the quality of MT are workshops like the one that was organized in conjunction with the MT Summit XIV in Nice, France, in the autumn of 2013 on "Multiword Units in Machine Translation and Translation Technology".[3] This effort, as well as other similar ones, tries to learn from all the theoretical work on MWEs to date, especially the work which has focused on different formalisms and techniques relevant for MWE processing in MT. These techniques include automatic recognition of MWEs in monolingual or bilingual corpora, alignment and paraphrasing methodologies, development and evaluation of handcrafted monolingual and bilingual linguistic resources and grammars, and use of MWEs in domain adaptation. Nonetheless, MWE translation issues have not yet been solved in a satisfactory manner, and there is still considerable room for improvement in all MT approaches whether knowledge-based, empirical (phrase-based, factored, syntax-based), or hybrid.

In general, it has been acknowledged that it is not possible to create large-scale NLP applications without a proper treatment of MWEs (Jackendoff, 1997; Zhang et al., 2006). But when it comes to the goals of the current project, we need to go a step further than current research in MWEs and deal with the semantic processing and representation of such expressions.

First, it is important to note that there is no uniform way to treat MWEs. MWEs occurring in a given text genre often require different treatment than MWEs occurring in another genre. Therefore, when handling MWEs, we will pay special attention to compositionality. The different degrees of compositionality influence the treatment MWEs, especially within deep grammars of natural language like the ones to be used in this

[2] http://clic2.cimec.unitn.it/starsem2013/

[3] http://mtsummit2013.info/workshop4.asp

project. For example, more compositional MWEs can be handled with grammar rules which allow for a flexible combination of the head word of such an MWE with other relevant lexemes (e.g., as in the approach described in Villavicencio et al. (2007)). Fixed expressions, on the other hand, can be added as word-with-spaces to the grammar lexicon.

We will therefore need a way to determine the degree of compositionality of a given MWE before attempting to integrate it into the grammar. This can be done by using crowdsourcing techniques and/or expert annotations in order to collect human judgements. Then, we can use these judgements to extract features deemed to be significant for determining the degree of compositionality and build a system which can automatically make graded judgements for expressions which are identified to be MWEs. Making finer distinctions with regard to compositionality of MWEs will help integrate better MWEs into the deep grammar, ultimately leading to better MT quality.

Once MWEs are properly handled by the deep grammar, we can make use of the formal semantic representations produced by the grammar and integrate MWEs into the MT system developed in the project. Since different types of grammars are used within the project, the produced semantic representation must be standardized. One possibility is to convert MRS and Prague-style representations into dependencies. Another form of common semantic representation is to use Linked Open Data as a sense inventory.

The output of the monolingual parsing should have ideally indicated an MWE in a sentence which is passed to the MT system. However, because of translation asymmetries, it may be the case that the transfer component of the system still splits the MWE and translates its components separately. To prevent such cases, we will focus on the identification and marking of MWEs in the transfer component of the MT system. Such marking will "force" the system to treat MWEs as single units which are not to be split and translated separately. We may use marked MWEs and their translations in the parallel training data in order to improve the transfer rules. Additionally, additional transfer rules can be learnt and new phrase pairs of MWEs can be aligned by using online resources such as Wiktionary[4] which contain lists of MWEs together with translations of those expressions in other languages. We may also benefit from the results of the TTC (`www.ttc-project.eu`) and ACCURAT (`www.accurat-project.eu`) projects, which deal with the production of comparable corpora for MT.

We will start pilot work with English, German, and Dutch, which will gradually be expanding to the other languages involved in the project as the other partners provide us with the necessary data. This will lead to improvement in MWEs identification and (deep) analysis of multiword expressions, in comparison with state-of-the-art from the dependency parsing. In close cooperation with the relevant partners, we will also transform the obtained parsing output into a format which is suitable for use in the MT prototypes and systems developed in QTLeap. The report on a first pilot of enhanced deep language processing systems will be reported by the end of month 12 of the project, as stated in the Description of Work of QTLeap. A second report will be prepared by the end of month 24. Those reports will be prepared in close cooperation with the other partners in Work Package 4 since the reports are concerned with improving the robustness of deep processing systems as a whole. This includes improved treatment of MWEs as well as improved semantic processing and natural language generation.

[4] `http://www.wiktionary.org/`

7 Expected Benefits for the Real Usage Scenario

The work done within the project on improving the quality of MT will be showcased in a real-life usage scenario. In this scenario, clients make requests or pose questions to an IT-helpdesk. These data are provided by the project partner HF and include short sentences, usually a request for help followed by an answer. In such a QA scenario the questions and answers are quite short, but even in such cases deep semantics would be useful, since the failure to analyse and translate the data correctly would lead to total miscommunication. Deliverable D4.1. provides clear examples in support of this claim.

Here, we concentrate on translation faults caused by incorrect treatment of MWEs. There is a considerable number of sentences in the data which contain various types of MWEs, phrasal verbs being the most common type. Below, we show some example sentences from the baseline translations produced in Pilot 0 which clearly demonstrate that MT in the real usage scenario can benefit significantly form the special treatment of MWEs proposed in QTLeap. The baseline Pilot 0 was created and applied on a subset of the HF data. The Moses translation systems was employed with a basic phrase-based setup (see deliverable D2.2). When reviewing the output sentences for both translation directions, some serious errors can be found that might be solvable by a good treatment of MWEs.
Consider the following English to German translation:

EN input sentence: *How do I get rid of this option?*
Pilot 0 DE output: *Wie bekomme ich rid diese Option?*

The MWE *get rid of* is not correctly translated in German. The MT system translates *get* to its literal German counterpart *bekommen* and the English word *rid* is not translated at all. As a result, the meaning of the German sentence cannot be understood at all. If the MT system had recognised *get rid of* as an MWE, it would have probably kept it intact in the translation table.

The next example shows an incorrect English to Bulgarian translation of a phrasal verb which leads to a wrong interpretation of the Bulgarian sentence. Bulgarian does not exhibit phrasal verbs in the same way English does. For example, an English phrasal verb is often translated into a single Bulgarian verb which further complicates the task of the MT system. For convenience, the Cyrillic is transcribed with Latin letters.

EN input sentence: *Is it possible to Configure Notepad++ to keep auto-completion always on?*
Pilot 0 BG output: *Vazmozhno li e da konfigurirate Notepad++ da zapazjat auto-completion vinagi v?*

The Bulgarian translation has numerous problems caused by wrong inflection. However, concentrating on the phrasal verb *keep on*, we note that it has not been recognised by the MT system as such which leads to a completely misleading interpretation in Bulgarian. The verb *keep* is translated in Bulgarian as *zapazjat* (to save, to preserve) and the particle *on* is just kept at the end of the sentence and translated as *v* (in). As a result, the Bulgarian sentence might be interpreted in the following way:

Is it possible to configure Notepad++ to save always in auto-completion?

Clearly, special treatment of MWE would probably have indicated that *keep on* is a phrasal verb and a better translation could have been produced.

Finally, as stated in Section 3, most of the existing techniques for treating MWEs will not perform well for MWEs with fully flexible structure. However, as it can be seen from the data analysis and the examples presented in this section, the vast majority of multiword lexical units encountered in the technical domain of the QTLeap data are phrasal verbs and compounds with relatively fixed structure which makes them easier to handle.

8 Conclusion

The current deliverable gives a general introduction to the theoretical aspects of MWEs as those have been identified and studied throughout the years. Without understanding those aspects, it is not possible to develop plausible and good quality computational techniques for handling MWEs in NLP applications.

We then summarised the general research trends in identifying MWEs in context and extracting them from corpora. In spite of the considerable progress made in those areas, MWE identification and extraction is not yet a solved problem. There is still room for the development of new techniques, focusing on sophisticated machine learning models like conditional random fields and spectral clustering.

We further emphasised on the fact that the integration of techniques to identify and interpret MWE into NLP applications improves the performance of those applications considerably. Given the goals of the project, we focused on recent research on MWE treatment within statistical MT.

Finally, goals for further improvement of the analysis of MWEs and their incorporation into MT have been set. The most important of those goals is to move towards a robust, adequate treatment of the semantics of MWEs and towards better MT transfer for such expressions.

Bibliography

Acosta, O. C., Villavicencio, A., and Moreira, V. P. (2011). Identification and treatment of multiword expressions applied to information retrieval. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011)*, pages 101–109, Portland, Oregon.

Baldwin, T. (2005). The deep lexical acquisition of english verb-particles. *Computer Speech & Language, Special Issue on Multiword Expressions*, 19(4):398–414.

Baldwin, T., Bender, E., Flickinger, D., Kim, A., and Oepen, S. (2004). Road-testing the English Resource Grammar over the British National Corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 2047–2050, Lisbon, Portugal.

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

Benor, S. B. and Levy, R. (2006). The chicken or the egg? A probabilistic analysis of english binomials. *Language*, 82:233–278.

Birke, J. and Sarkar, A. (2006). A clustering approach for the nearly unsupervised recognition of nonliteral language. In *Proceedings of the 11th Conference of the EACL (EACL 2006)*, pages 329–336, Trento, Italy.

Blunsom, P. (2007). *Structured Classication for Multilingual Natural Language Processing.* PhD thesis, University of Melbourne.

Carpuat, M. and Diab, M. (2010). Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.*, HLT '10., pages 242–245, Stroudsburg, PA, USA. Association for Computational Linguistics.

Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Constant, M., Roux, J. L., and Sigogne, A. (2013). Combining compound recognition and PCFG-LA parsing with word lattices and conditional random fields. *ACM Transactions on Speech and Language Processing (TSLP) - Special issue on multiword expressions: From theory to practice and use*, 10(3).

Copestake, A. and Flickinger, D. (2000). An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the 2nd International Conference on Language Resource and Evaluation (LREC 2000)*, Athens, Greece.

Copestake, A. A., Lambeau, F., Villavicencio, A., Bond, F., Baldwin, T., Sag, I. A., and Flickinger, D. (2002). Multiword expressions: linguistic precision and reusability. In *LREC*.

Cruse, A. D. (1986). *Lexical Semantics*. Cambridge University Press.

de Medeiros, H., Ramisch, C., Volpe, M., and Villavicencio, A. (2010). Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 44:59–77.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Fazly, P., Cook, P., and Stevenson, S. (2009). Unsupervised type and token identification of idiomatic expressions. *Comput. Linguistics*, 35(1):61–103.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. The MIT press.

Fillmore, C., Kay, P., and O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions. *Language*, 64:501–538.

Firth, J. R. (1957). *Papers in Linguistics 1934–1951*. Oxford University Press.

Gurrutxaga, A. and Alegria, I. (2012). Measuring the compositionality of nv expressions in basque by means of distributional similarity techniques. In *Proceedings of LREC 2012*.

Hashimoto, C. and Kawahara, D. (2008). Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 992–1001, Honolulu, Hawaii.

Jackendoff, R. (1997). *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.

Katz, G. and Giesbrecht, E. (2006). Automatic identification of non-compositional multiword expressions using latent semantic analysis. In *Proceedings of the ACL-2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 28–35, Sydney, Australia.

Kim, S. N. and Baldwin, T. (2010). How to pick out token instances of english verb-particle constructions. In *Journal of Language Resources and Evaluation (LRE) : Special Issue on Multiword Expressions: hard going or plain sailing?*, pages 97–113. Language Resources and Evaluation.

Lapata, M. and Lascarides, A. (2003). Detecting novel compounds: The role of distributional evidence. In *Proceedings of the 11th Conference of the European Chapter for the Association of Computational Linguistics (EACL-2003)*, pages 235–242, Budapest, Hungary.

Lau, J. H., Baldwin, T., and Newman, D. (2013). On collocations and topic models. *ACM Transactions on Speech and Language Processing (TSLP) - Special issue on multiword expressions: From theory to practice and use*, 10(2).

Lin, D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 317–324, College Park, USA.

McCarthy, D., Keller, B., and Carroll, J. (2003). Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, Sapporo, Japan.

Morin, E. and Daille, B. (2010). Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation Special Issue on Multiword expression: hard going or plain sailing.*, pages 79–95.

Nunberg, G., Sag, I. A., and Wasow, T. (1994). Idioms. *Language*, 70:491–538.

Pal, S., Naskar, S. K., Pecina, P., Bandyopadhyay, S., and Way, A. (2010). Handling named entities and compound verbs in phrase-based statistical machine translation. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 46–54. Coling 2010 Organizing Committee.

Patrick, J. and Fletcher, J. (2005). Classifying verb particle constructions by verb arguments. In *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 200–209, Colchester, UK.

Pearce, D. (2001). Synonymy in collocation extraction. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pages 41–46, Pittsburgh, USA.

Pecina, P. (2008). *Lexical Association Measures*. PhD thesis, Charles University Prague.

Ramisch, C., Schreiner, P., Idiart, M., and Villavicencio, A. (2008). An evaluation of methods for the extraction of multiword expressions. In Grégoire, N., Evert, S., and Krenn, B., editors, *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 50–53.

Ramisch, C., Villavicencio, A., and Kordoni, V., editors (2013). *ACM Transactions on Speech and Language Processing (TSLP) - Special issue on multiword expressions: From theory to practice and use*, volume 10.

Rayson, P., Piao, S., Sharoff, S., Eevert, S., and Moiron, B. V., editors (2010). *Special Issue on Multiword Expression: Hard Going or Plain Sailing*, volume 44. Springer.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing.*, CICLing '02., pages 1–15, London, UK. Springer-Verlag.

Schone, P. and Jurafsky, D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, pages 100–108, Hong Kong, China.

Simova, I. and Kordoni, V. (2013). Improving English-Bulgarian statistical machine translation by phrasal verb treatment. In *Proceedings of MT Summit XIV Workshop on Multi-word Units in Machine Translation and Translation Technology*, Nice, France.

Sporleder, C. and Li, L. (2009). Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the EACL (EACL 2009)*, pages 754–762, Athens, Greece.

van der Beek, L. (2005). The extraction of determinerless pp. In *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 190–199, Colchester, UK.

van Noord, G. (2004). Error mining for wide-coverage grammar engineering. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*, pages 446–453, Barcelona, Spain.

Villavicencio, A., Bond, F., Korhonen, A., and McCarthy, D., editors (2005). *Computer Speech & Language (Special issue on Multiword Expressions)*, volume 19. Elsevier.

Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M., and Ramisch, C. (2007). Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1034–1043, Prague, Czech Republic.

Zhang, Y., Kordoni, V., Villavicencio, A., and Idiart, M. (2006). Automated multiword expression prediction for grammar engineering. In *Proceedings of the ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 36–44, Sydney, Australia.