This is a pre-print of an article published in the ACM TSLP journal.
The final authenticated version is available online at: https://doi.org/10.1145/2483691.2483692

Introduction to the Special Issue on Multiword Expressons: From Theory to Practice and Use

CARLOS RAMISCH, Université Joseph Fourier
ALINE VILLAVICENCIO, Federal University of Rio Grande do Sul
VALIA KORDONI, Humboldt Universität zu Berlin

We are in 2013, and multiword expressions have been around for a while in the computational linguistics research community. Much has been discussed, proposed, experimented, evaluated and argued about them in the last 12 years, since the first ACL workshop on MWEs in Sapporo, Japan. And yet, they deserve the publication of a whole special issue of the ACM Transactions on Speech and Language Processing. But what is it about multiword expressions that makes them never get out of fashion? Today, who are the people and the institutions who perform and publish groundbreaking fundamental and applied research in this field? What is the place and the relevance of our lively research community in the bigger picture of computational linguistics? Where do we come from as a community, and most importantly, where are we heading to? In this introduction paper, we share our point of view about the answers to these questions and introduce the papers that compose the current special issue.

ACM Reference Format:
Carlos Ramisch, Aline Villavicencio and Valia Kordini. 2013. Multiword Expressons: From Theory to Practice and Use — Introduction to the Special Issue *ACM Trans. Speech Lang. Process.* V, N, Article A (January YYYY), 10 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

As guest editors of the special issue on multiword expressions of the ACM Transactions on Speech and Language Processing, we proudly present a selection of high-quality research papers on the subject. The articles selected to appear in the current journal issue constitute a body of original research that brings significant contributions to the field, shedding some light on challenging issues like semantic interpretation and parsing of multiword expressions. We have received a total of thirty-one submissions, from which we selected, with the help of the reviewers, a set of nine papers, yielding a competitive acceptance rate of 29%. For all the hard work, we are grateful to the authors, reviewers, the TSLP editorial board and staff.

In this introductory paper, we draw a panorama of the current research carried out in our community, from our point of view as guest editors. Therefore, we start by presenting briefly the history and the main challenges that need to be tackled in MWE treatment. We discuss some relevant seminal papers on the acquisition, interpretation, representation and application of multiword expressions in computational linguistics. Afterwards, we summarise the contributions published in this special issue and point out the current initiatives like the annual MWE workshop and the recent foundation of a dedicated SIGLEX Section on MWEs. We conclude on our view of the place of our field in computational linguistics and discuss some possibilities of future directions.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© YYYY ACM 1550-4875/YYYY/01-ARTA $15.00
DOI: http://dx.doi.org/10.1145/0000000.0000000

This is a pre-print of an article published in the ACM TSLP journal.
The final authenticated version is available online at: https://doi.org/10.1145/2483691.2483692

2. WHERE DO WE COME FROM?

The study of MWEs is almost as old as linguistics itself. When trying to classify linguistic phenomena into either lexical, syntactic, semantic, and so on, one realises that some of them, and in particular MWEs, lie in between two levels. As a consequence, any linguistic and computational approach to language with some level of semantic interpretation has to include MWE representations in its models, in order to achieve robustness.

It is widely assumed that the number of MWEs in any human language is of the same order of magnitude than the number of simple words [Jackendoff 1997]. They occur frequently in all language registers, from written to oral, from general to domain-specific, etc. The English Wordnet, for example, contains only a limited amount of relatively fixed MWEs, but still, out of its 117,827 nouns, 60,292 (51.4%) are multiword; and out of its 11,558 verbs, 2,829 (25.5%) are multiword. Unfortunately, the English Wordnet can be seen as an exception as far as lexical resources are considered. Because of their heterogeneous properties, MWEs are hard to acquire and to model, and therefore are often ignored in the construction of computational lexicons.

If, on the one hand, MWEs represent a challenge for computational linguistics because of their complex behavior, on the other hand their treatment represents an important goal, in order to deal with a wide range of linguistic phenomena that are often a weak point in many NLP applications. MWEs are nowadays a hot topic and an exciting area of computational linguistics. Research has made significant progress in recent years, and this is reflected by the large number of papers that focus on data-driven (semi-)automatic acquisition of multiword units from corpora. A considerable body of techniques, resources and tools to perform automatic extraction from texts is now available, and we discuss the increase in interest in the topic in the next sections.[1]

2.1. On consensus and diversity

In Figure 1, we plot the ratio of papers in the ACL anthology which mention "multiword", "collocation" or "idiom" with respect to the total number of papers in the years 1965 to 2006. This is an evidence of the growing importance of the automatic MWE acquisition field within computational linguistics.

The graphic also shows that there are multiple ways of referring to MWEs, and that their use varies over time, with *multiword expressions* being the most popular term since 2000, probably because of the workshop series and other such initiatives. Because of the history of the field, researchers with different linguistic backgrounds name MWEs in various ways, with slightly different meanings and connotations. Some prefer to call them *collocations*, emphasizing the tendency that words have to co-occur. Others call them *compounds* or *complex words*, as in compound nouns and complex predicates. In the lexicographic tradition, they are often referred to as *multiword units* or *polylexical expressions*. *Phraseology* [Mel'čuk et al. 1995] is another term employed to describe recurrent word combinations or *formulaic language* [Biber et al. 1999; Wray 2002], specially in technical and scientific domains [Cabré 1992]. In construction grammar they are called *idioms*, underlining their lack of semantic compositionality [Fillmore et al. 1988]. The term *multiword expression* was made popular by the Stanford project and by the famous "pain-in-the-neck" paper [Sag et al. 2002]. All these terms refer to similar yet slightly different notions, each one underlining a different property of MWEs.

The truth is that, we must admit, we do not even agree on the orthography for multiword or multi-word expression! To start with, there is no single definition of

[1] See http://multiword.sf.net/ for a non exhaustive list.

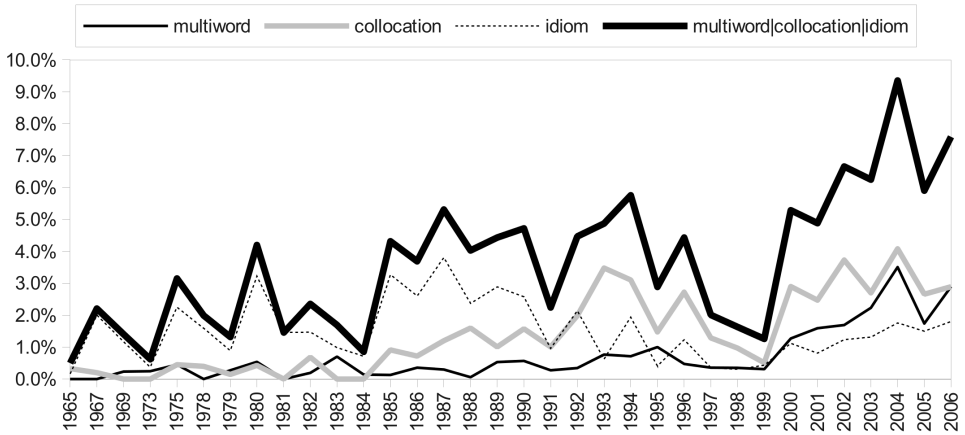

Fig. 1. Ratio of papers mentioning "multiword", "collocation" or "idiom" over the total number of papers published in a year in the ACL Anthology. Proportions queried from the ACL-ACR corpus version 20090501, available at http://acl-arc.comp.nus.edu.sg/.

what actually a MWE is and what it is not. Something which probably every research would agree on is that there is more to language than atomic elements being combined through compositional rules. In that sense, probably Baldwin and Kim [2010] provide the most generic definition, stating that "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".

Even though there are different views on MWEs, this does not fragment the research community. On the contrary, these views reflect the relevance of MWEs for several distinct areas of NLP, given that MWEs are recurrent and pervasive constructions not only in human languages, but also (and as a consequence) in computational linguistics. Indeed, they are often cited as challenges for future work in papers on machine translation, word sense disambiguation and parsing. As a consequence, researchers from all these areas have a common goal: to better take into account MWEs, improving the quality of the results generated by NLP applications. In short, diversity of backgrounds enriches the MWE research community, making it extremely lively and interesting.

2.2. A historical overview

In computational linguistics, the study of MWEs arose from the availability of very large corpora and of computers capable of analysing them by the end of the 80's and beginning of the 90's. One of the main goals of these first attempts to process MWEs using machines was to build systems for computer-assisted lexicography and terminography of multiword units. Among the seminal papers of the field, one of the most often cited is by Choueka [1988], who proposed a method for collocation extraction based on n-gram statistics.

Another groundbreaking work is that of Smadja [1993]. He proposed Xtract, a tool for collocation extraction based on some simple POS filters and on mean and standard deviation of word distance. His approach had the advantage of handling non-contiguous constructions. It was strongly based on the notion of collocation as outstanding co-occurrence.

Church and Hanks [1990] suggested the use of a more sophisticated association measure based on mutual information. They provided theoretical justification for it and then tested it on relatively large corpora for the extraction of terminological and collocational units. Later, Dagan and Church [1994] proposed a terminographic environment called Termight, which used this association score. Termight performed bilingual extraction and provided tools to easily classify candidate terms, find bilingual correspondences, define nested terms and investigate occurrences through a concordancer.

Also in the context of terminographic extraction, Justeson and Katz [1995] proposed a simple approach based on a small set of POS patterns and frequency thresholds. Using minimal linguistic information combined with an intuitive quantitative technique, they obtained surprisingly good results given the simplicity of the technique.[2]

The indiscriminate use of association measures was criticised by Dunning [1993]. He argued that the assumption underlying most measures is that words are distributed normally, but corpus evidence dos not support this hypothesis. Therefore, he proposed a 2-gram measure called *likelihood ratio*, that estimates directly how more likely a 2-gram is than expected by chance. In addition to being theoretically sound, Dunning's score is also easily interpretable. Nowadays, measures based on likelihood ratio (e.g., the log-likelihood score) are still largely employed in several MWE extraction contexts.

At the beginning of the 2000's, the Stanford MWE project[3] has revived interest of the NLP community in this topic. One of the most cited publications of the MWE project is the famous "pain-in-the-neck" paper by Sag et al. [2002]. It provided an overview of MWE characteristics and types and then presented some methods for dealing with them in the context of grammar engineering. The Stanford MWE project is also at the origin of the MWE workshop series, which started in 2001 and are in their 9th edition in 2013[4].

2.3. The MWE community

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. Nowadays, the MWE research community is organised in the form of a Section of SIGLEX. 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 [Bond et al. 2003; Tanaka et al. 2004; Rayson et al. 2006; Moirón et al. 2006; Grégoire et al. 2007; Grégoire et al. 2008; Anastasiou et al. 2009; Laporte et al. 2010; Kordoni et al. 2011]. 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.[5] The site also hosts a repository with several annotated data sets 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 workshop series, the *SEM conference[6] features a track on MWEs and all main com-

[2] A pedagogical example of the application of this technique on a corpus is given by [Manning and Schütze 1999, p. 156].
[3] http://mwe.stanford.edu/
[4] http://multiword.sourceforge.net/mwe2013
[5] http://multiword.sourceforge.net/
[6] http://clic2.cimec.unitn.it/starsem2013/

putational linguistics conferences such as COLING, ACL and LREC regularly feature papers on MWEs.

Some reference book chapters and textbooks are available for researchers wanting to know more about MWEs. In one of the main NLP textbooks, Manning and Schütze [1999] provide, in Chapter 5, a detailed introduction on collocation extraction using association measures, including worked out examples. The book chapter by McKeown and Radev [1999] summarises some linguistic aspects and more advanced results focusing on collocations. Another book chapter is the one by Baldwin and Kim [2010], who present a generic and structured view of the field and summarise a broad range of the relevant literature on MWE acquisition and interpretation. In her book, Seretan [2011] discusses how different linguistic theories deal with collocations and presents experimental work on the use of deep parsing for their acquisition. Several Ph.D. theses were also published in the area and provide deeper insights into the challenges of MWE processing.

Researchers and research groups all over the world perform fundamental and applied research in MWE treatment. For example, since 2010, 91 people have reviewed for the MWE workshops and special issue. They are affiliated to 81 institutions in 25 countries, with the most represented countries being the UK, Germany, the USA, France, Japan, Canada, the Netherlands and Brazil. Another example of collective effort is the PARSEME network, an European COST Action where participants investigate the interaction between parsing and MWEs.[7]

As a complement to workshops and conferences, this is the third special issue published by leading journals in computational linguistics. The first was the journal of Computer Speech and Language [Villavicencio et al. 2005] and the second was the journal of Language Resources and Evaluation [Rayson et al. 2010]. 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.

3. HOW FAR HAVE WE GOT?

The papers included in this special issue provide a cross-section of the current research in the area, highlighting some of the main topics and proposing a variety of approaches for MWE tasks, from their identification to their interpretation. To set the tone for the special issue an invited contribution by Kenneth Church, *How Many Multiword Expressions do People Know?*, steps back and discusses some of the relevant questions that have been driving research on MWEs and examining how they are addressed in a variety of fields such as linguistics, lexicography, educational testing and web search.

The remaining contributions exemplify the various tasks involved in the life-cycle of MWEs, and give a sampler of the particular challenges posed by different types of MWEs.

In the first of these, *Lexical Semantic Factors in the Acceptability of English Support-Verb-Nominalization Constructions*, Anthony Davis and Leslie Barrett look at the properties of support-verb and nominalization in English, like *take a walk* and *give a talk*, examining some linguistic factors for potential correlations with the acceptability of these constructions. In particular, they investigate whether acceptability of support verb and nominalization pairs is linked to membership in Levin's verbal classes [Levin 1993], evaluating around 2,700 (acceptable and unacceptable) combinations.

Veronika Vincze, István Nagy T. and János Zsibrita examine the task of identification in *Learning to Detect English and Hungarian Light Verb Constructions*, using a tool based on conditional random fields, the FXTagger. They look at domain specificity

[7] http://www.cost.eu/domains_actions/ict/Actions/IC1207

This is a pre-print of an article published in the ACM TSLP journal.
The final authenticated version is available online at: https://doi.org/10.1145/2483691.2483692

of LVCs and check the portability of the models generated from different corpora, using domain adaptation techniques to reduce the gap between these domains.

In *Modelling the internal variability of multiword expressions through a pattern-based method*, Malvina Nissim and Andrea Zaninello look at ways of optimising, that is, improving recall without losing precision, in a flexible search for the token identification of Italian MWEs in corpora. They propose a method for modelling the internal variation of MWEs, in terms of frequent variation patterns of specific part-of-speech sequences, focusing on two types of nominal expressions. When compared against association measures, these variation patterns produced more precise information.

In addition to identifying MWEs, a related challenge lies in incorporating them in NLP tasks like parsing, word sense disambiguation and topic modelling, and this is the focus of the next three papers.

Combining Compound Recognition and PCFG-LA Parsing with Word Lattices and Conditional Random Fields by Matthieu Constant, Joseph Le Roux and Anthony Sigogne, examines MWEs in parsing. They 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 specialised annotation schemes for compounds. These solutions bring improvements that are often reflected in parsing accuracy.

In *Word Sense and Semantic Relations in Noun Compounds*, Su Nam Kim and Timothy Baldwin focus on word sense disambiguation of noun compounds, proposing a method for disambiguating the senses of the component nouns, and using the predicted senses in an interpretation task. They argue that, even if the component words are polysemous in isolation, when they occur as part of an noun compound they display sense preferences. Thus, the confines of the compound narrow down the possible senses, e.g. *plant* in isolation with 3 senses and disambiguated to one sense in *coffee plant*. To disambiguate word senses, they built WSD classifiers obtaining an accuracy of around 55% and 37% respectively in supervised and in unsupervised setups, outperforming a benchmark WSD system.

On Collocations and Topic Models, by Jey Han Lau, Timothy Baldwin and David Newman, discusses the impact of handling MWEs and adding them to the document representation for topic modelling. The prior identification of MWEs and their treatment as single tokens creates more parsimonious models and improves topic coherence. The model fit is measured using the Akaike information criterion, which penalises model complexity. They find improvements in topic quality when MWEs are explicitly incorporated.

MWEs present a wide range of idiomaticity, ranging from more transparent and compositional combinations such as *cheese knife* and *salt and pepper* to idiomatic cases like *trip the light fantastic* and *make ends meet*. Considerable effort has been devoted to determining whether words are used literally or figuratively in a given context, and to what extent they contribute to the meaning of a given MWE. This is important for a variety of NLP tasks and applications that perform some degree of semantic analysis, such as machine translation, question answering, opinion mining and information retrieval, to ensure precision and naturalness of the results.

The semantic interpretation of MWEs is discussed by several of the papers in the special issue, in particular by Ekaterina Shutova, Jakub Kaplan, Simone Teufel and Anna Korhonen in *A Computational Model of Logical Metonymy*. Logical metonymy occurs in predicates like *enjoy the book* vs *enjoy the cake* for *enjoy reading the book* and *enjoy eating the cake*. The authors propose a model that combines statistical techniques and linguistic theory for the interpretation of metonymic phrases. This produces sense-level interpretations for metonymic phrases and groups them into conceptual classes, achieving an F-measure of 0.64 compared to the human agreement 0.76.

This is a pre-print of an article published in the ACM TSLP journal.
The final authenticated version is available online at: https://doi.org/10.1145/2483691.2483692

The relevance of handling MWEs can be assessed in terms of their impact in applications which involve semantic interpretation. For example, treating an idiomatic MWE as a semantic unit can prevent misunderstandings and loss of information, as is the case of *having a riot* as *having a good time* with a positive profile for sentiment analysis, and translating *Big Apple* as *New York* for an MT system. This is the topic of the next two contributions.

The paper by Beata Beigman Klebanov, Jill Burstein and Nitin Madnani, *Sentiment Profiles of Multi-Word Expressions in Test-Taker Essays: The Case of Noun-Noun Compounds*, proposes an approach to measure the compositionality of the sentiment profile of an MWE and applies it to noun-noun compounds. For instance, an expression like *heart attack* has a strongly negative sentiment profile due to the strongly negative profile of the head word. The authors also examine the impact of using the sentiment profiles of compounds in a sentiment classification system, and find improvements in performance of sentence-level sentiment polarity classification on both essay and product reviews data.

Preslav Nakov and Marti Hearst focus on MT, presenting an approach for the interpretation of noun compounds in *Semantic Interpretation of Noun Compounds Using Verbal and Other Paraphrases*. In particular, they investigate the use of paraphrases with predicates making explicit the relation between the component words of a noun compound. For instance, for *migraine treatment* such predicates would convey the noun compound as referring to a treatment for *relieving* or *preventing* migraine. They propose an approach for finding paraphrasing verbs and prepositions for noun compounds, and investigate the use of the paraphrasing verbs with weights for the representation of their semantics. They apply the verb-based explicit paraphrases in the context of statistical MT, obtaining an improvement in translation quality.

4. WHERE DO WE HEAD TO?

As briefly mentioned in section 2.3, the constantly growing scientific community researching MWEs has very recently been further organised in a section of SIGLEX, the Special Interest Group on the Lexicon of the Association for Computational Linguistics. This is a very strong indication that research specialising in multiword expressions has reached out to attract the attention of the broader computational linguistics and language technology research community, as this has been formed under the umbrella of the ACL. This is a very good platform for the development of MWE research, as well as for the research community in the future.

As far as this future is concerned, let us try to give our insights here by splitting it into a near future and a bit more distant one.

For the near future, and on the basis of current and emerging trends, the focus clearly lies in applications which seek to incorporate MWEs and benefit from this incorporation in efficiency.

First of all, research in statistical machine translation (SMT) has already turned its interest to MWEs. Specifically, phrase-based SMT systems which rely solely on mappings on the level of word sequences may produce inadequate translations in cases where complex linguistic phenomena occur, like some types of MWEs. For accurate, precise and natural high quality translation results, SMT has acknowledged that it is important to incorporate an adequate treatment for MWEs, whose interpretation poses a challenge given their idiosyncratic, flexible and heterogeneous nature. Some 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 word-with-spaces, and a dynamic strategy, that keeps a record of the number of MWEs in the source phrase. They found that both strategies result

This is a pre-print of an article published in the ACM TSLP journal.
The final authenticated version is available online at: https://doi.org/10.1145/2483691.2483692

in improvement of translation quality, which suggests that SMT phrases alone may not model all MWE information. Pal et al. [2010] also found improvements brought by applying preprocessing steps like single-tokenization along with prior alignment and transliteration for named entities and compound verbs. Morin and Daille [2010] obtained 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 (e.g. *chronic fatigue syndrome* decomposed as *[chronic fatigue] [syndrome]*, *[chronic] [fatigue syndrome]* or *[chronic] [fatigue] [syndrome]*).

Characteristic of this trend of research on exploitation of MWEs for improvement of the quality of machine translation are workshops like the one to be organised in conjunction with the MT Summit XIV in Nice, France in the autumn of 2013 on "Multiword Units in Machine Translation and Translation Technology".[8] 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 hand-crafted 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 proper treatment of MWEs.

But what is it exactly that we should be looking into when talking about the more distant future in the research of MWEs? In one word, the answer would be semantics, of course.

The semantics of a MWE may influence the performance of NLP tasks. In parsing, for instance, the identification of more idiomatic phrasal verbs, which subcategorisation properties may differ from those of the simplex verbs, is not problematic (for statistical parsers in general). However, highly compositional cases (e.g., *call in*) may be less distinct syntactically from verb-PP combinations, requiring additional information, such as selectional preferences [Kim and Baldwin 2010].

The impact, though, that the semantics and the semantic compositionality of a MWE may have on the quality of the outcome of real-life applications has not been researched thoroughly in the literature to date. And this is what future research on MWEs should set as one of its first priorities. Adequate treatment of the semantics of MWEs will not only help the correct interpretation of MWE in applications, but it will also enable their full integration into the underlying models.

Specifically, the general idea is to make use of formal representations related to the semantics of MWEs and its analysis in order to help integrate MWEs in already existing systems for (S)MT, IE, IR, etc., where various kinds of deep and shallow linguistic and extra-linguistic information is at work. That is, the future research needs to deal with the semantics of MWEs and produce semantic representations from deep and shallow resources, aiming for a maximal degree of compatibility. The most important reason to take this approach is that work on providing a standardised semantic representation would be useful for many purposes: allowing different tools in real-life applications, like parsers and generators, for instance, to be connected up to a variety of underlying systems. Standardising grammatical relations is not directly useful in this way. Furthermore, semantic representations constructed from shallow tools em-

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

bedded in NLP systems can be viewed as underspecified forms of the output from a deeper tool in a way that can be made formally precise. Finally, any constraints that arise from the domain are most naturally expressed in terms of semantics and it is potentially very useful to be able to apply domain constraints cross-domain, as well as cross-framework.

The current issue represents a snapshot of an important moment for the MWE research community. On the one hand, MWE identification and extraction is not yet a solved problem. There is still room for the development of new techniques for MWE acquisition, focusing on sophisticated machine learning models like conditional random fields and spectral clustering. On the other hand, MWE treatment in NLP applications is emphasised, with articles that cover a varied range of environments like educational testing, sentiment analysis, machine translation, topic modelling, word sense disambiguation and parsing. Some articles also shed some light on linguistic models and computational processing of complex semantic interpretation tasks like the acceptability of support verbs and the compositionality of metonymy. Thus, the current issue constitutes a significant step forward towards the goals of the near future of MWE research. We hope that the next editions of the MWE workshop and possibly a next special issue on MWEs features more papers on these subjects, thus contributing to an even stronger consolidation and increase in visibility for our community.

Acknowledgements

We would like to thank the support of LICIA and of projects CAPES/COFECUB Cameleon (707/11), CNPq 551964/2011-1, 482520/2012-4 and 478222/2011-4.

REFERENCES

Dimitra Anastasiou, Chikara Hashimoto, Preslav Nakov, and Su Nam Kim (Eds.). 2009. *Proc. of the ACL Workshop on MWEs: Identification, Interpretation, Disambiguation, Applications (MWE 2009)* (6). ACL, Suntec, Singapore. http://aclweb.org/anthology-new/W/W09/W09-29 70 p.

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In *Handbook of Natural Language Processing* (2 ed.), Nitin Indurkhya and Fred J. Damerau (Eds.). CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 267–292.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English* (1st ed.). Pearson Education Ltd, Harlow, Essex. 1204 p.

Francis Bond, Anna Korhonen, Diana McCarthy, and Aline Villavicencio (Eds.). 2003. *Proc. of the ACL Workshop on MWEs: Analysis, Acquisition and Treatment (MWE 2003)* (12). ACL, Sapporo, Japan. http://aclweb.org/anthology-new/W/W03/W03-1800 104 p.

Maria Teresa Cabré. 1992. *La terminologia. La teoria, els mètodes, les aplicacions*. Empúries, Barcelona, Spain. 527 p.

Marine Carpuat and Mona Diab. 2010. Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation. In *Proc. of HLT: The 2010 Annual Conf. of the NAACL (NAACL 2003)*. ACL, Los Angeles, California, 242–245. http://www.aclweb.org/anthology/N10-1029

Yaacov Choueka. 1988. Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in Large Textual Databases. In *Proceedings of the 2nd International Conference on Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications - RIA 1988)* (21-25), Christian Fluhr and Donald E. Walker (Eds.). CID, Cambridge, MA, USA, 609–624.

Kenneth Church and Patrick Hanks. 1990. Word Association Norms Mutual Information, and Lexicography. *Comp. Ling.* 16, 1 (1990), 22–29.

Ido Dagan and Kenneth Church. 1994. Termight: Identifying and Translating Technical Terminology. In *Proc. of the 4th ANLP Conf. (ANLP 1994)*. ACL, Stuttgart, Germany, 34–40. DOI:http://dx.doi.org/10.3115/974358.974367

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Comp. Ling.* 19, 1 (1993), 61–74.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. *Language* 64 (Sep. 1988), 501–538. Issue 3. http://www.jstor.org/stable/414531

Nicole Grégoire, Stefan Evert, and Su Nam Kim (Eds.). 2007. *Proc. of the ACL Workshop on A Broader Perspective on MWEs (MWE 2007)*. ACL, Prague, Czech Republic. http://aclweb.org/anthology-new/W/W07/W07-11 80 p.

Nicole Grégoire, Stefan Evert, and Brigitte Krenn (Eds.). 2008. *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)* (1). Marrakech, Morocco. http://www.lrec-conf.org/proceedings/lrec2008/workshops/W20_Proceedings.pdf 57 p.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. Number 28 in Linguistic Inquiry Monographs. MIT Press, Cambridge, MA, USA. 262 p.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Nat. Lang. Eng.* 1, 1 (1995), 9–27.

Su Nam Kim and Timothy Baldwin. 2010. How to pick out token instances of English verb-particle constructions. *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing* 44, 1-2 (Apr. 2010), 97–113. DOI:http://dx.doi.org/10.1007/s10579-009-9099-7

Valia Kordoni, Carlos Ramisch, and Aline Villavicencio (Eds.). 2011. *Proc.of the ACL Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011)* (23). ACL, Portland, OR, USA. http://www.aclweb.org/anthology/W/W11/W11-08 144 p.

Éric Laporte, Preslav Nakov, Carlos Ramisch, and Aline Villavicencio (Eds.). 2010. *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)* (28). ACL, Beijing, China. http://aclweb.org/anthology-new/W/W10/W10-37 89 p.

Beth Levin. 1993. *English Verb Classes and Alternations: a preliminary investigation*. Chicago UP, Chicago, USA.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, USA. 620 p.

Kathleen R. McKeown and Dragomir R. Radev. 1999. Collocations. In *A Handbook of Natural Language Processing*, Robert Dale, Hermann Moisl, and Harold Somers (Eds.). Marcel Dekker, New York, NY, USA, Chapter 15, 507–523.

Igor Mel'čuk, André Clas, and Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Editions Duculot, Louvain la Neuve, Belgium. 256 p.

Begoña Villada Moirón, Aline Villavicencio, Diana McCarthy, Stefan Evert, and Suzanne Stevenson (Eds.). 2006. *Proc. of the COLING/ACL Workshop on MWEs: Identifying and Exploiting Underlying Properties (MWE 2006)* (23). ACL, Sidney, Australia. http://aclweb.org/anthology-new/W/W06/W06-12 61 p.

Emmanuel Morin and Béatrice Daille. 2010. Compositionality and lexical alignment of multi-word terms. *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing* 44, 1-2 (Apr. 2010), 79–95. DOI:http://dx.doi.org/10.1007/s10579-009-9098-8

Santanu Pal, Sudip Kumar Naskar, Pavel Pecina, Sivaji Bandyopadhyay, and Andy Way. 2010. Handling Named Entities and Compound Verbs in Phrase-Based Statistical Machine Translation. In *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)* (28), Éric Laporte, Preslav Nakov, Carlos Ramisch, and Aline Villavicencio (Eds.). ACL, Beijing, China, 45–53.

Paul Rayson, Scott Piao, Serge Sharoff, Stefan Evert, and Begoña Villada Moirón (Eds.). 2010. *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing*. Vol. 44. Springer.

Paul Rayson, Serge Sharoff, and Svenja Adolphs (Eds.). 2006. *Proc. of the EACL Workshop on MWEsin Multilingual Context (EACL-MWE 2006)* (3). ACL, Trento, Italy. http://aclweb.org/anthology-new/W/W06/W06-2400 79 p.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proc. of the 3rd CICLing (CICLing-2002)* (17-23) *(LNCS)*, Vol. 2276/2010. Springer, Mexico City, Mexico, 1–15.

Violeta Seretan. 2011. *Syntax-Based Collocation Extraction* (1st ed.). Text, Speech and Language Technology, Vol. 44. Springer, Dordrecht, Netherlands. 212 p.

Frank A. Smadja. 1993. Retrieving Collocations from Text: Xtract. *Comp. Ling.* 19, 1 (1993), 143–177.

Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen (Eds.). 2004. *Proc. of the ACL Workshop on MWEs: Integrating Processing (MWE 2004)* (26). ACL, Barcelona, Spain. http://aclweb.org/anthology-new/W/W04/W04-0400 103 p.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy (Eds.). 2005. *Comp. Speech & Lang. Special issue on MWEs*. Vol. 19. Elsevier.

Alison Wray. 2002. *Formulaic Language and the Lexicon*. Cambridge UP, Cambridge, UK. 348 p.